

EXHIBIT 22



Lennart Johnsson
2015-02-06

The Impact of Moore's law and loss of Dennard Scaling

Lennart Johnsson
University of Houston



Outline

The quest for Energy Efficiency in Computing

- Technology changes and the impacts thereof
- A retrospective on benefits and drawbacks of past architectural changes
- Expected architectural changes and impact
- Our research to understand opportunities and challenges for HPC driven by the expected technology changes



Lennart Johnsson
2015-02-06

What got me interested in energy efficient computing

Energy cost estimate for a ~1300 node cluster purchase 2008 for PDC @ Royal Institute of Technology:

Four year energy and cooling cost
~1.5 times cluster cost incl. software,
maintenance and operations!!



Foto: Harald Barth



Business as usual not appealing from a
scientist/user point of view



Lennart Johnsson

2015-02-06

“You Can Hide the Latency,
But,

You Cannot Hide the ENERGY!!”

Peter M Kogge

Rule of thumb $1\text{MW} = \$1\text{M}/\text{yr}$
Average US household $\sim 11\text{MWh}/\text{yr}$
($1\text{MW yr} \approx 800$ households)

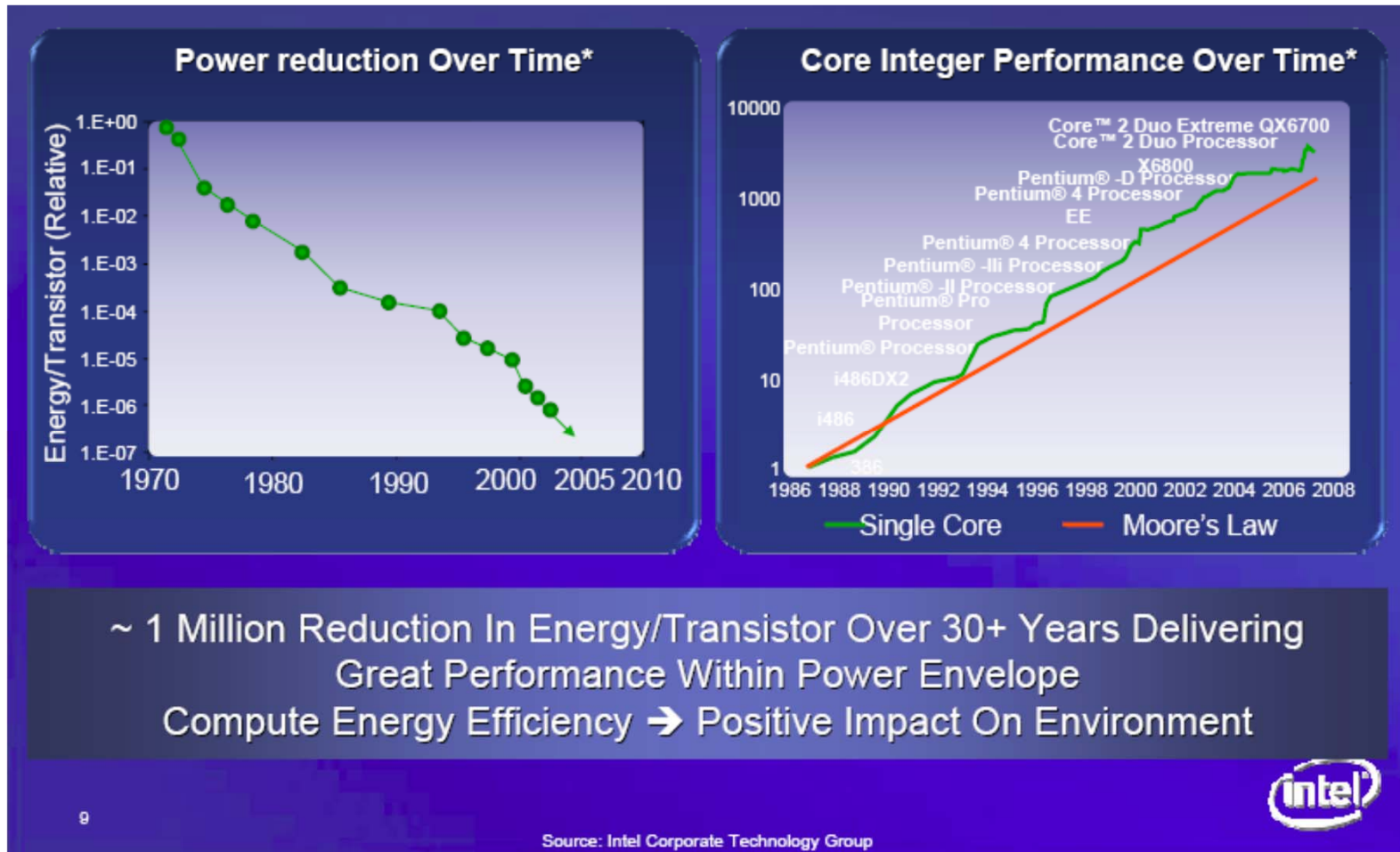


UNIVERSITY of HOUSTON



Lennart Johnsson
2015-02-06

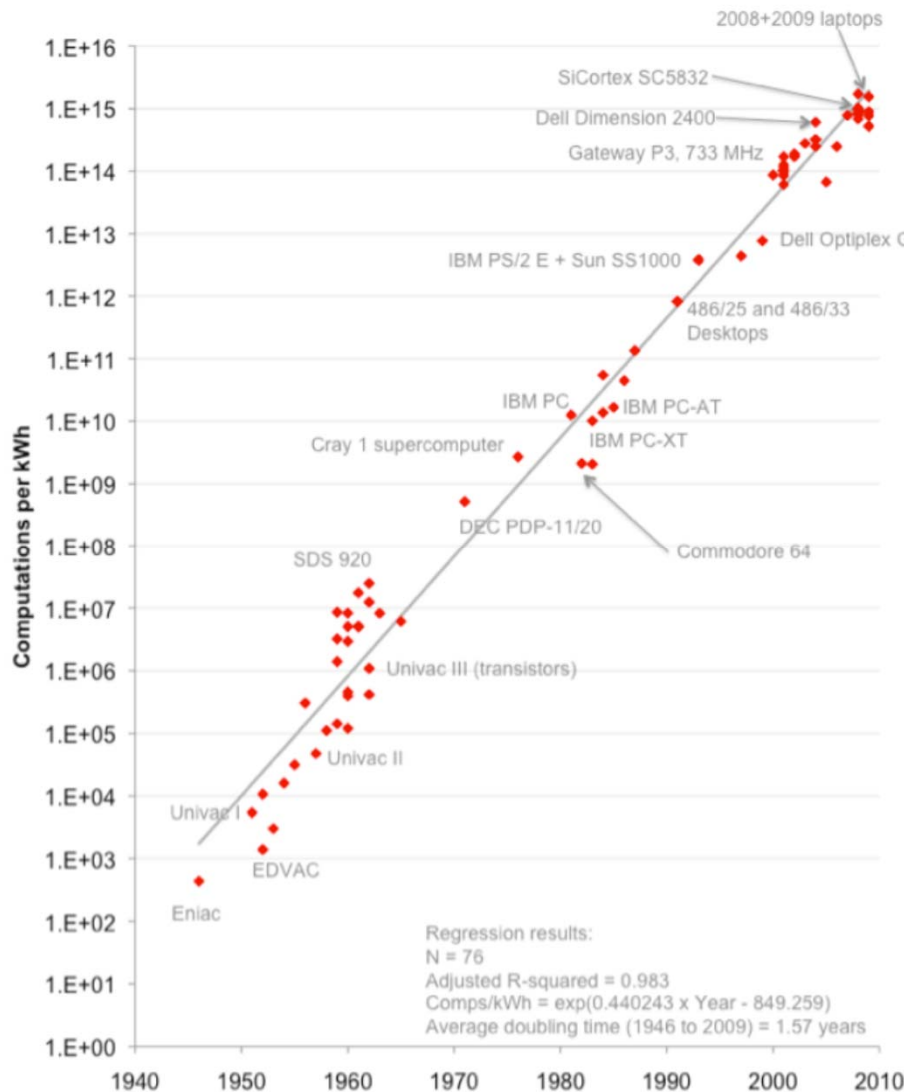
Incredible Improvement in Integrated Circuit Energy Efficiency



Source: Lorie Wigle, Intel, http://piee.stanford.edu/cgi-bin/docs/behavior/becc/2008/presentations/18-4C-01-Eco-Technology_-_Delivering_Efficiency_and_Innovation.pdf

UNIVERSITY of HOUSTON

Energy efficiency evolution



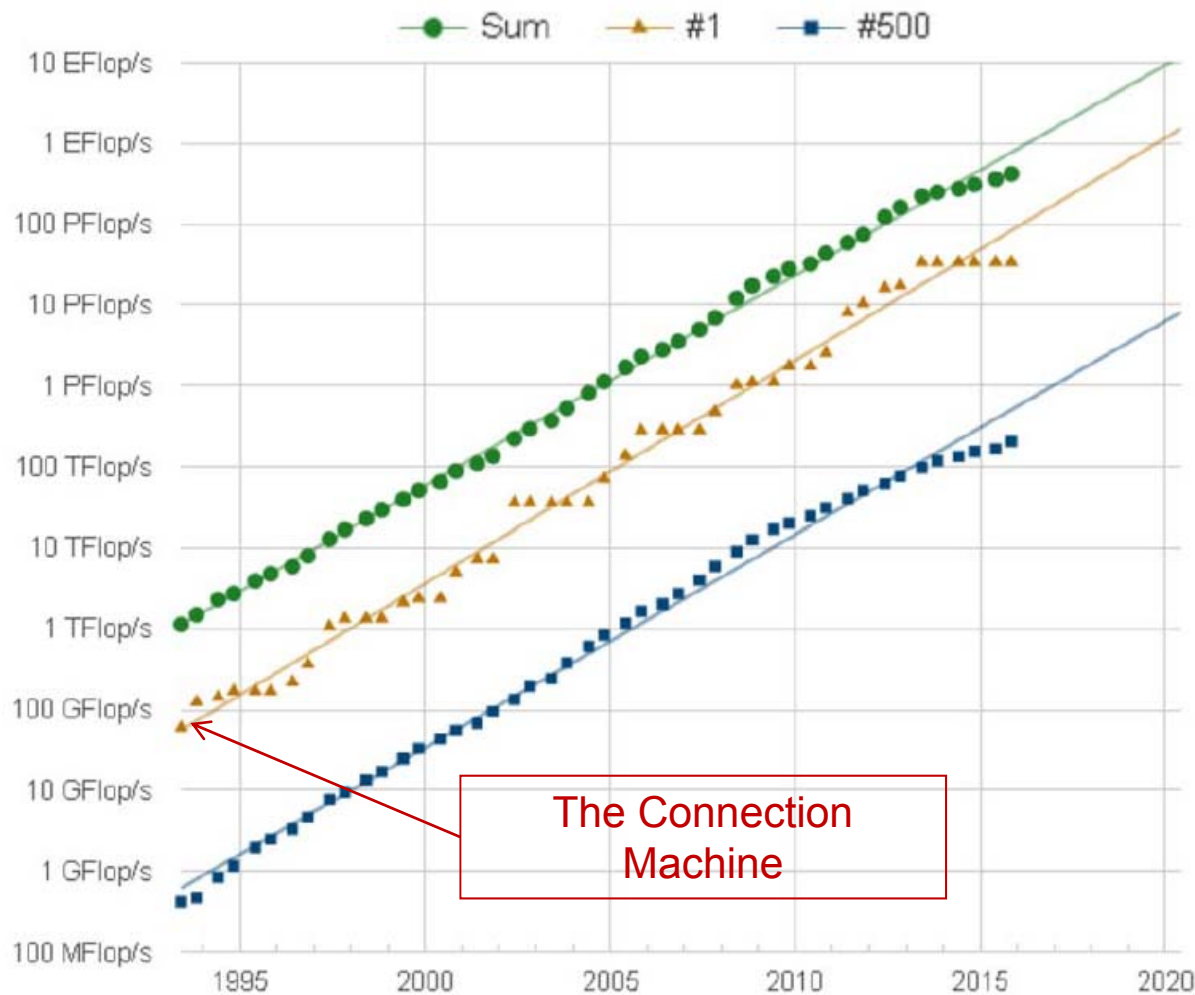
Energy efficiency
doubling every 18.84
months on average
measured as
computation/kWh

Source: Assessing in the Trends in the
Electrical Efficiency of Computation over
Time,
J.G. Koomey, S. Berard, M. Sanchez, H.
Wong, Intel, August 17, 2009,
[http://download.intel.com/pressroom/pdf/
computertrendsrelease.pdf](http://download.intel.com/pressroom/pdf/computertrendsrelease.pdf)



Lennart Johnsson
2015-02-06

Top500 system performance evolution



Performance doubling
period on average:

No 1 – 13.64
months

No 500 – 12.90
months

www.top500.org

UNIVERSITY of HOUSTON

WSOU-ARISTA001466

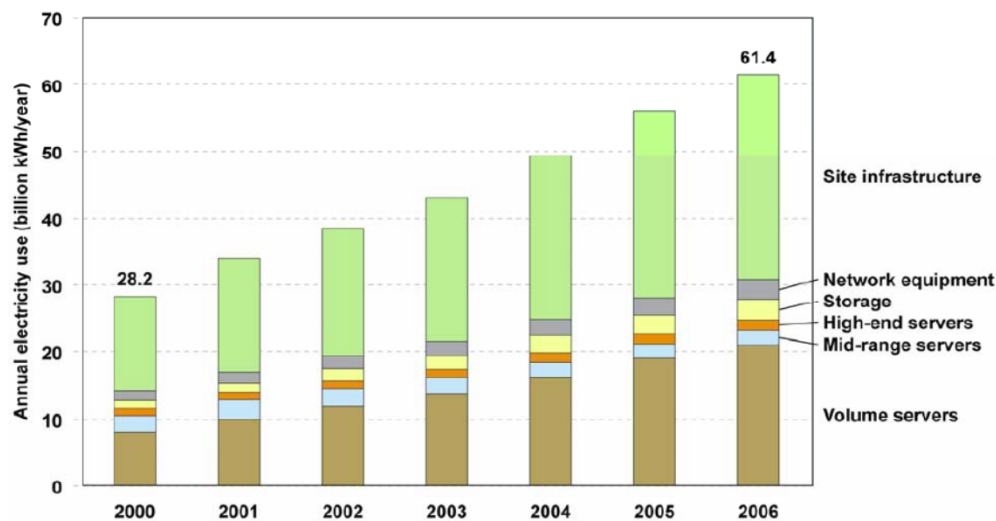


Lennart Johnsson
2015-02-06

The Gap

The energy efficiency improvement as determined by Koomey does not match the performance growth of HPC systems as measured by the Top500 list

The Gap indicates a growth rate in energy consumption for HPC systems of about 20%/yr.



EPA study projections: 14% - 17%/yr

Uptime Institute projections: 20%/yr

PDC experience: 20%/yr

Report to Congress on Server and Data Center Energy Efficiency”,
Public Law 109-431, U.S Environmental Protection Agency,
Energy Star Program, August 2, 2007,
http://www.energystar.gov/ia/partners/prod_development/download/s/EPA_Datacenter_Report_Congress_Final1.pdf

End use component	2000		2006		2000 – 2006 electricity use CAGR
	Electricity use (billion kWh)	% Total	Electricity use (billion kWh)	% Total	
Site infrastructure	14.1	50%	30.7	50%	14%
Network equipment	1.4	5%	3.0	5%	14%
Storage	1.1	4%	3.2	5%	20%
High-end servers	1.1	4%	1.5	2%	5%
Mid-range servers	2.5	9%	2.2	4%	-2%
Volume servers	8.0	29%	20.9	34%	17%
Total	28.2		61.4		14%

“Findings on Data Center Energy Consumption Growth May
Already Exceed EPA’s Prediction Through 2010!”,
K. G. Brill, The Uptime Institute, 2008,
<http://uptimeinstitute.org/content/view/155/147>



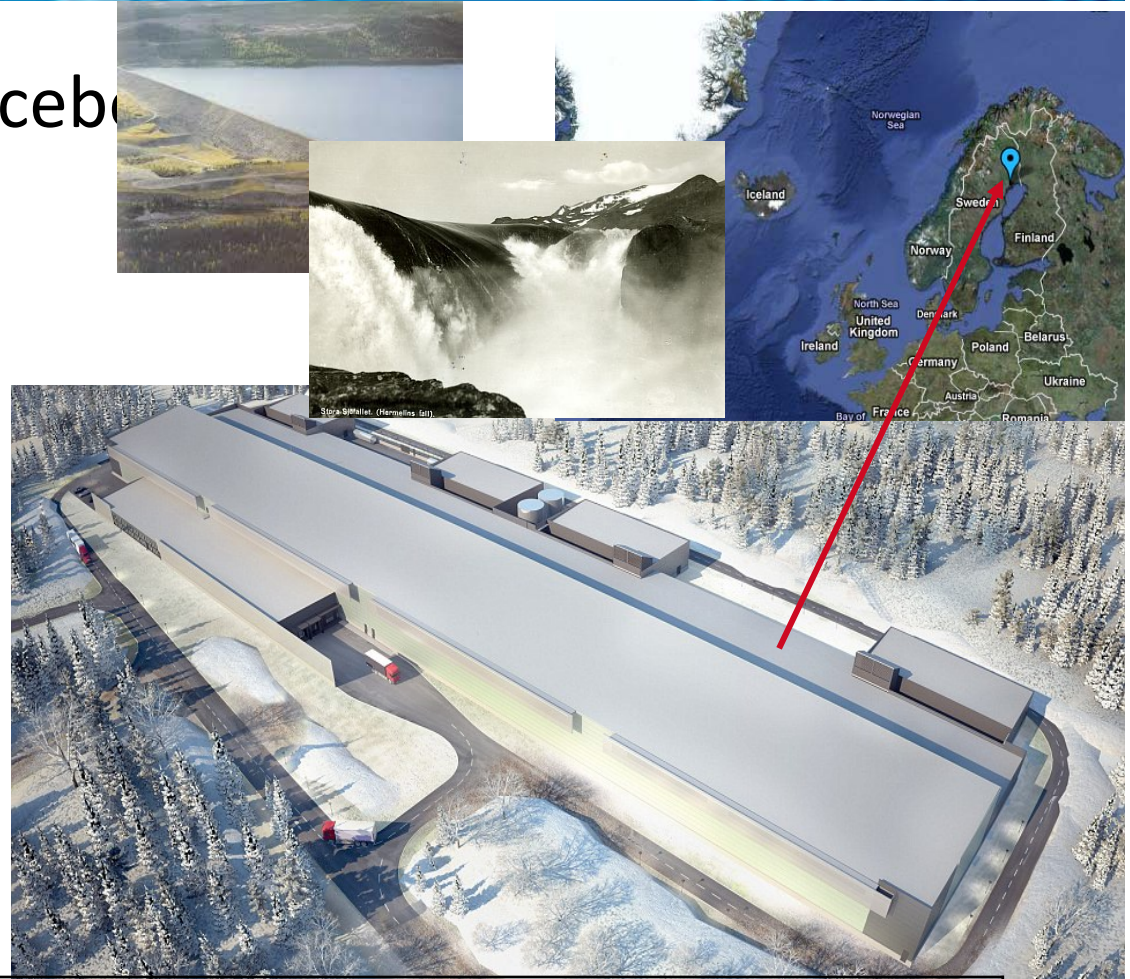
Lennart Johnsson
2015-02-06

Internet Data Centers -Facebook

- The 120 MW Lulea Data Center will consist of three server buildings with an area of 28 000 m² (300 000 ft²). The first building is to be operational within a year and the entire facility is scheduled for completion by 2014
- The Lulea river has an installed hydroelectric capacity of 4.4 GW and produces on average 13.8 TWh/yr

Read more:

<http://www.dailymail.co.uk/sciencetech/article-2054168/Facebook-unveils-massive-data-center-Lulea-Sweden.html#ixzz1diMHIYIL>



Climate data for Luleå, Sweden

2011-10-28

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Average high ° C (° F)	-8 (18)	-8 (18)	-2 (28)	3 (37)	10 (50)	17 (63)	19 (66)	17 (63)	12 (54)	6 (43)	-1 (30)	-5 (23)	5.0 (41.0)
Average low ° C (° F)	-16 (3)	-16 (3)	-11 (12)	-4 (25)	2 (36)	8 (46)	11 (52)	10 (50)	5 (41)	1 (34)	-7 (19)	-13 (9)	-2.5 (27.5)

UNIVERSITY of HOUSTON



Actions of large consumers/providers

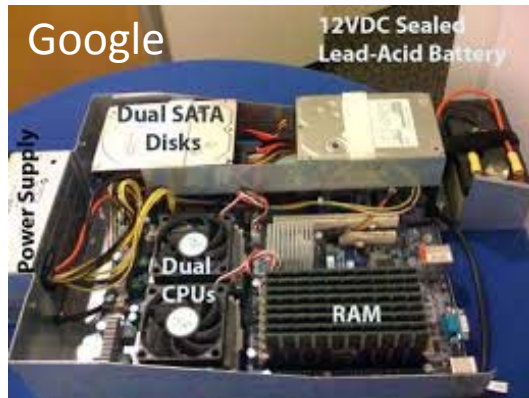
- Custom server designs for reduced energy consumption, stripping out un-needed components, and in some cases reducing redundancy implementing resiliency at the system level, not at the server level
- In some cases making designs public, e.g. Facebooks Open Compute initiative
- Focusing on clean renewable energy
- Locating data centers close to energy sources (low transmission losses) and where cooling costs are low





Lennart Johnson
2015-02-06

Internet Company effort examples



<http://www.epanorama.net/newepa/2009/05/07/google-efficient-data-centers/>



<http://arstechnica.com/information-technology/2013/02/who-needs-hp-and-dell-facebook-now-designs-all-its-own-servers/>

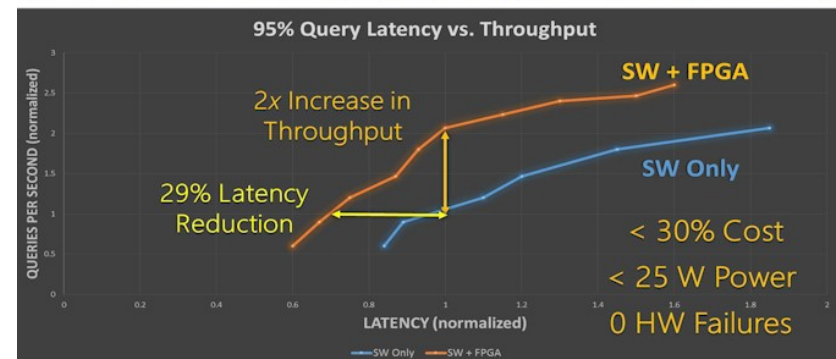
Microsoft

1,632 Servers with FPGAs Running Bing Page Ranking Service (~30,000 lines of C++)



- Two 8-core Xeon 2.1 GHz CPUs
- 64 GB DRAM
- 4 HDDs @ 2 TB, 2 SSDs @ 512 GB
- 10 Gb Ethernet
- No cable attachments to server

Air flow
200 LFM
68 °C Inlet



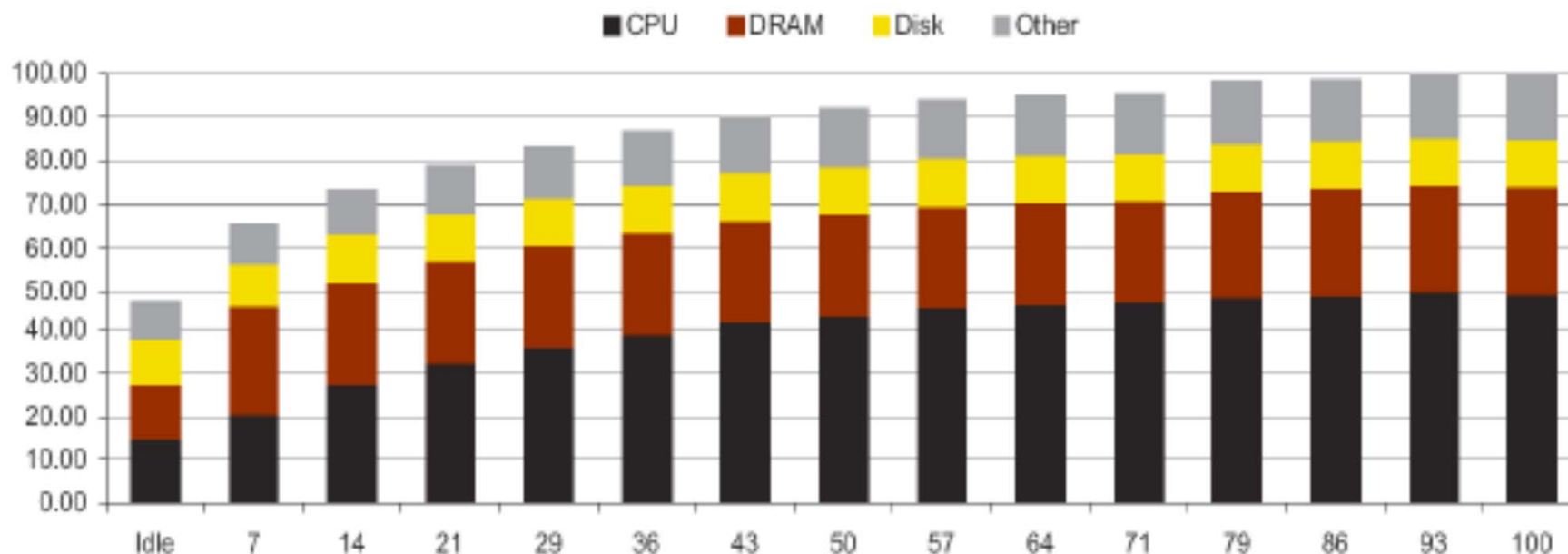
<http://www.enterprisetech.com/2014/09/03/microsoft-using-fpgas-speed-bing-search/>

UNIVERSITY of HOUSTON



Lennart Johnsson
2015-02-06

Power Consumption – Load for a typical server (2008/2009)



**CPU power consumption at low load about 40% of consumption at full load.
Power consumption of all other system components independent of load,
approximately.**

Result: Power consumption at low load about 65% of consumption at full load.

Luiz Andre Barroso, Urs Hoelzle, The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines <http://www.morganclaypool.com/doi/pdf/10.2200/s00193ed1v01y200905cac006>



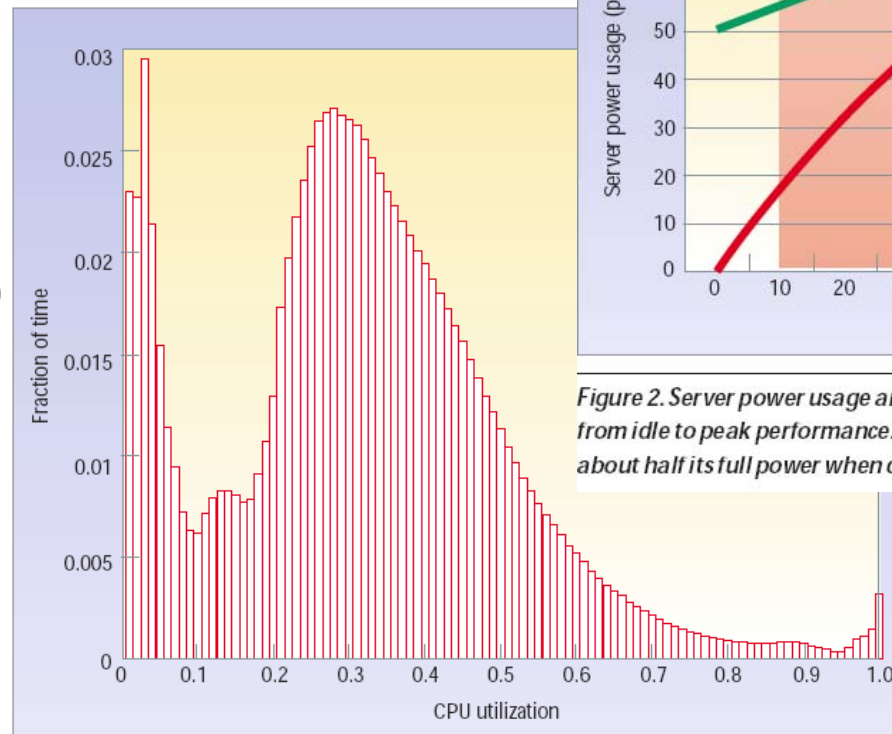
Lennart Johnsson
2015-02-06

COVER FEATURE

The Case for Energy-Proportional Computing

Luiz André Barroso and Urs Hölzle
Google

Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum utilization levels.



Google

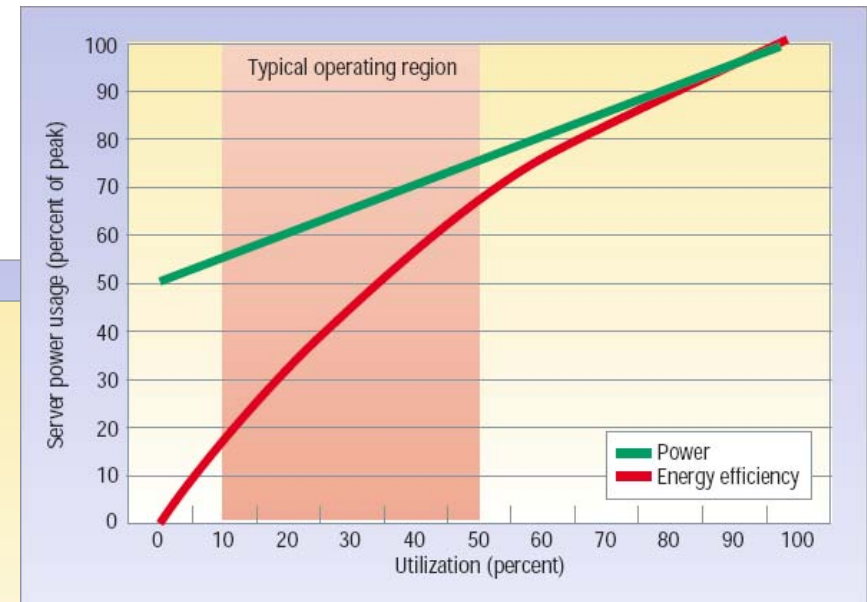


Figure 2. Server power usage and energy efficiency at varying utilization levels, from idle to peak performance. Even an energy-efficient server still consumes about half its full power when doing virtually no work.

“The Case for Energy-Proportional Computing”, Luiz André Barroso, Urs Hölzle, *IEEE Computer*, vol. 40 (2007).
http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//pubs/archive/33387.pdf



Lennart Johnsson
2015-02-06

Internet vs HPC Workloads

Google (Internet)

KTH/PDC (HPC)

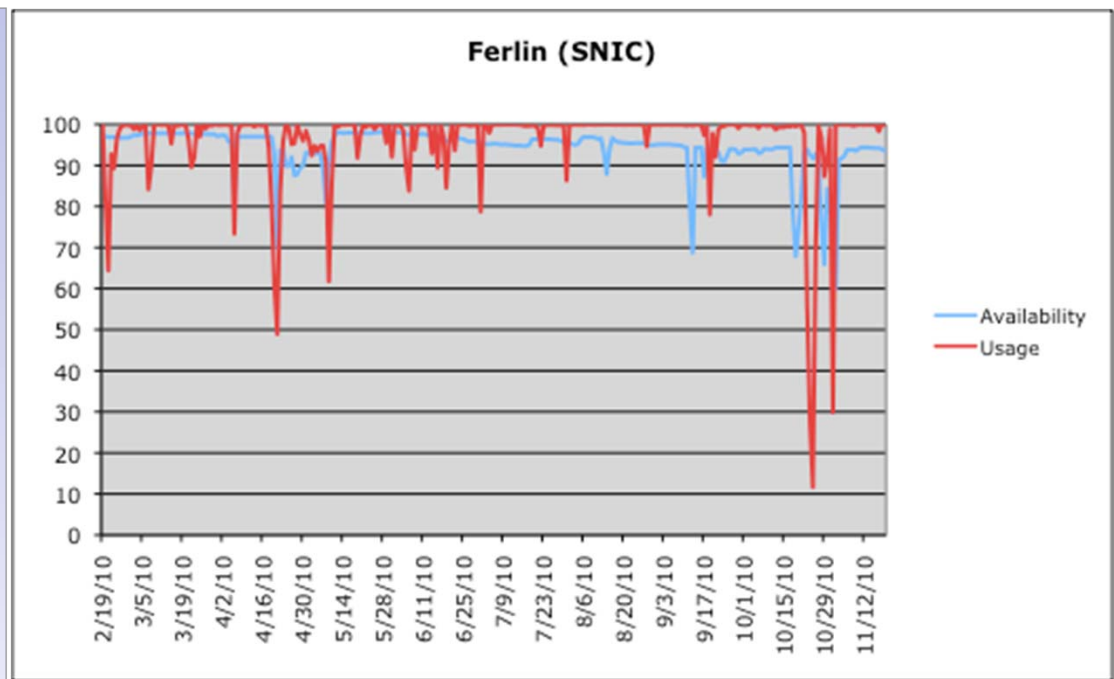
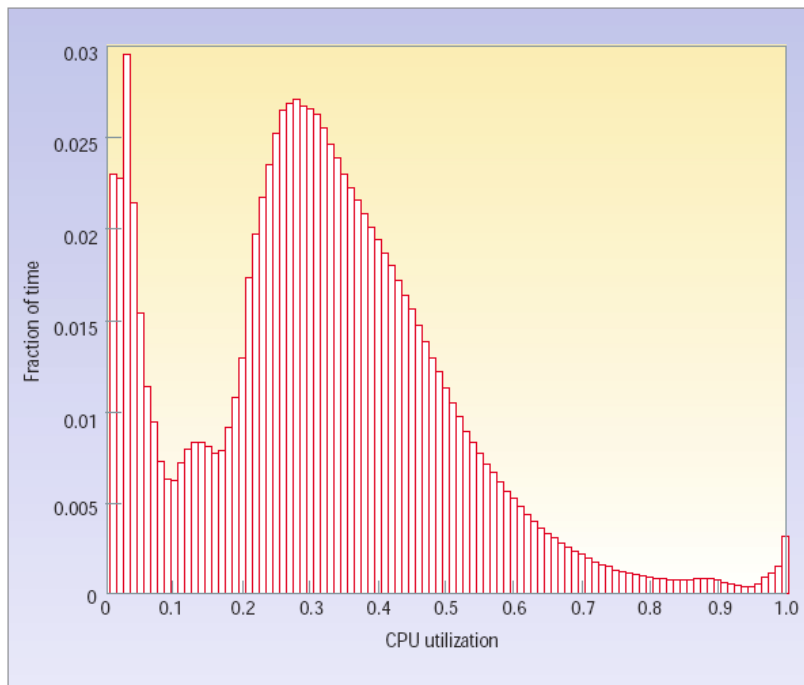


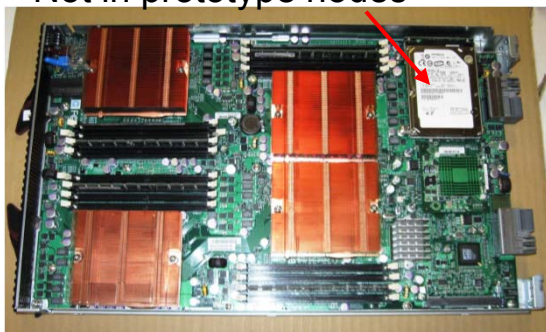
Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum utilization levels.



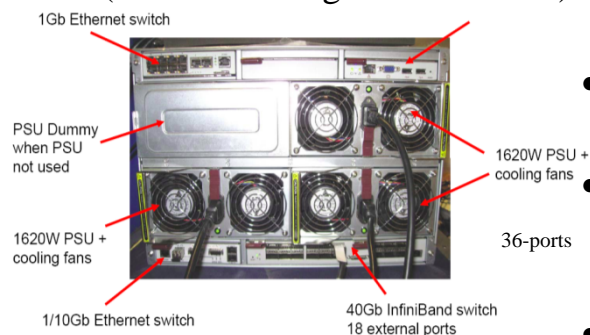
Lennart Johnsson
2015-02-06

The 1st foray into energy efficient computing “on a budget”: the SNIC/KTH/PRACE Prototype I

Not in prototype nodes

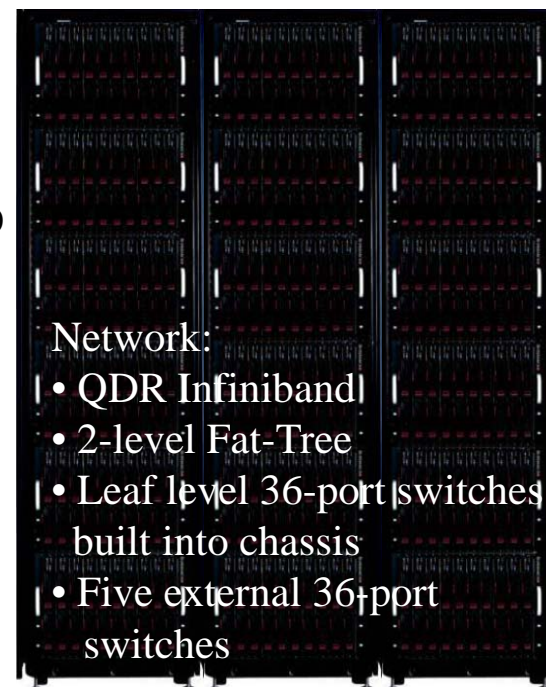


CMM (Chassis Management Module)



Objective: Energy efficiency at par with most energy efficient design at the time, but using only **commodity technology (for cost)** and **no acceleration for preservation of programming paradigm**

- New 4-socket blade with 4 DIMMs per socket supporting PCI-Express Gen 2 x16
- Four 6-core 2.1 GHz 55W ADP AMD Istanbul CPUs, 32GB/node
- 10-blade in a 7U chassis with 36-port QDR IB switch, new efficient power supplies.
- 2TF/chassis, 12 TF/rack, 30 kW (6 x 4.8)
- 180 nodes, 4320 cores, full bisection QDR IB interconnect



Network:

- QDR Infiniband
- 2-level Fat-Tree
- Leaf level 36-port switches built into chassis
- Five external 36-port switches



WSOU-ARISTA001474



The Prototype HPL Efficiencies in Perspective

Power

Dual socket Intel Nehalem 2.53 GHz	240 MF/W
Above + GPU	270 MF/W
PRACE/SNIC/KTH prototype	344 MF/W (unoptimized)
IBM BG/P	357 - 372 MF/W

Fraction of Peak

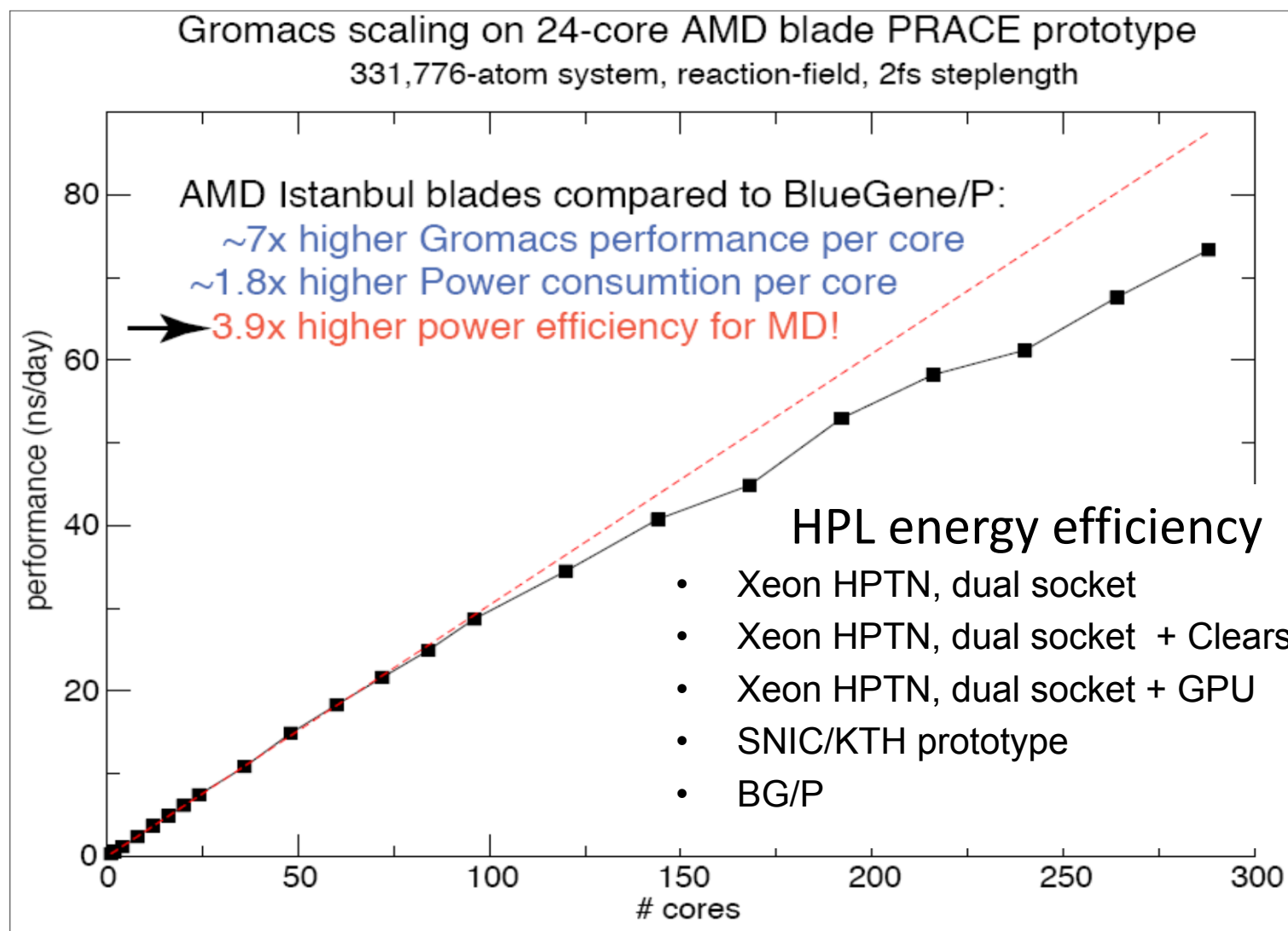
Dual socket Intel Nehalem 2.53 GHz	91%
Above + GPU	53%
PRACE/SNIC/KTH prototype	79%
IBM BG/P	83%



Lennart Johnsson

2015-02-06

Comparison with Best Proprietary System



UNIVERSITY of HOUSTON



Lennart Johnsson
2015-02-06



Reducing Waste

Mark Horowitz 2007: “Years of research in low-power embedded computing have shown only one design technique to reduce power: reduce waste.”



Seymour Cray 1977: “Don’t put anything in to a supercomputer that isn’t necessary.”



Exascale Computing Technology Challenges, John Shalf
National Energy Research Supercomputing Center, Lawrence Berkeley National Laboratory
ScicomP / SP-XXL 16, San Francisco, May 12, 2010

UNIVERSITY of HOUSTON



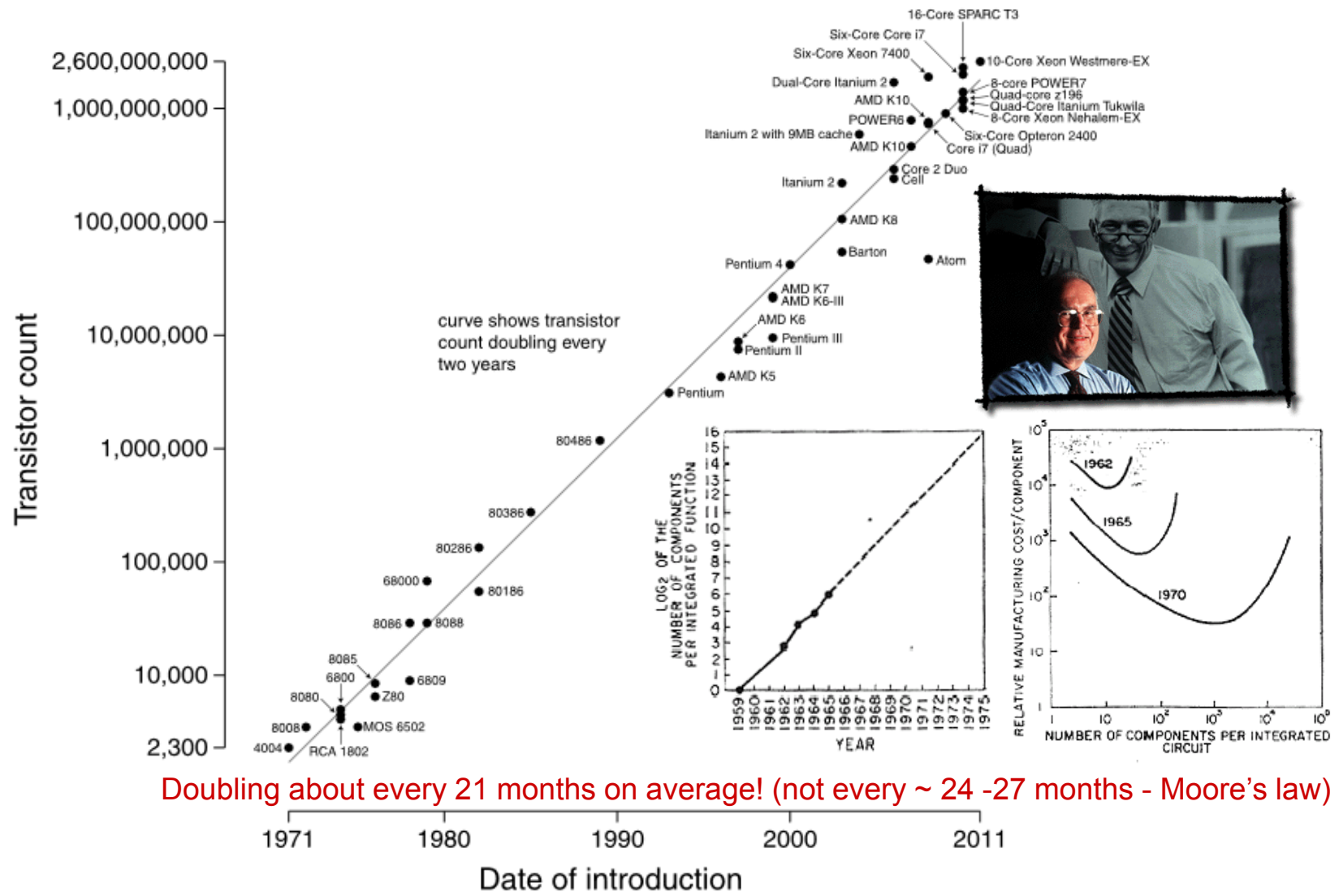
The Good News

- **Moore's Law still works** and expected to work throughout the decade
- The number of **transistors per die** (processor) on average **has increased slightly faster** than predicted by Moore's law (for several reasons, such as, slightly increased die sizes (on average), changing balance of transistors used for logic and memory, and evolving device technology)



Lennart Johnsson
2015-02-06

Microprocessor Transistor Counts 1971-2011 & Moore's Law



Doubling about every 21 months on average! (not every ~ 24 -27 months - Moore's law)

http://upload.wikimedia.org/wikipedia/commons/0/00/Transistor_Count_and_Moore%27s_Law_-_2011.svg

UNIVERSITY of HOUSTON

WSOU-ARISTA001479



Lennart Johnsson

2015-02-06

Moore's Law at work

Processor	Transistors	Year	Manufact.	Process	Area
Dual-Core Itanium 2	1,700,000,000	2006	Intel	90 nm	596 mm ²
POWER6	789,000,000	2007	IBM	65 nm	341 mm ²
Six-Core Opteron 2400	904,000,000	2009	AMD	45 nm	346 mm ²
RV870	2,154,000,000	2009	AMD	40 nm	334 mm ²
16-Core SPARC T3	1,000,000,000	2010	Sun/Oracle	40 nm	377 mm ²
Six-Core Core i7	1,170,000,000	2010	Intel	32 nm	240 mm ²
8-Core POWER7	1,200,000,000	2010	IBM	45 nm	567 mm ²
4-Core Itanium Tukwila	2,000,000,000	2010	Intel	65 nm	699 mm ²
8-Core Xeon Nehalem-EX	2,300,000,000	2010	Intel	45 nm	684 mm ²
Cayman (GPU)	2,640,000,000	2010	AMD	40 nm	389 mm ²
GF100 (GPU)	3,000,000,000	2010	nVidia	40 nm	529 mm ²
AMD Interlagos (16 C)	2,400,000,000	2011	AMD	32 nm	315 mm ²
AMD GCN Tahiti (GPU)	4,310,000,000	2011	AMD	28 nm	365 mm ²
10-Core Xeon Westmere-EX	2,600,000,000	2011	Intel	32 nm	512 mm ²
8-Core Itanium Poulson	3,100,000,000	2012	Intel	32 nm	544 mm ²
Sandy Bridge, 8C	2,270,000,000	2012	Intel	32 nm	435 mm ²
Ivy Bridge, 4C+GPU	1,400,000,000	2012	Intel	22 nm	160 mm ²
Ivy Bridge, 6C	1,860,000,000	2013	Intel	22 nm	257 mm ²
Ivy Bridge, 15C	4,300,000,000	2013	Intel	22 nm	541 mm ²
Vertex-7 (FPGA)	3,500,000,000	2013	Xilinx	28 nm	550 mm ² ?
Nvidia Kepler GK110 (GPU)	7,100,000,000	2013	nVidia	28 nm	551 mm ²
Xeon Phi, 62C	5,000,000,000	2013	Intel	22 nm	?

Doubling about every 21 months on average (not every ~ 24 -27 months!)

UNIVERSITY of HOUSTON



The Good News

- **Moore's Law still works** and expected to work throughout the decade
- The number of **transistors per die** (processor) on average **has increased slightly faster** than predicted by Moore's law (for several reasons, such as, slightly increased die sizes (on average), changing balance of transistors used for logic and memory, and evolving device technology)

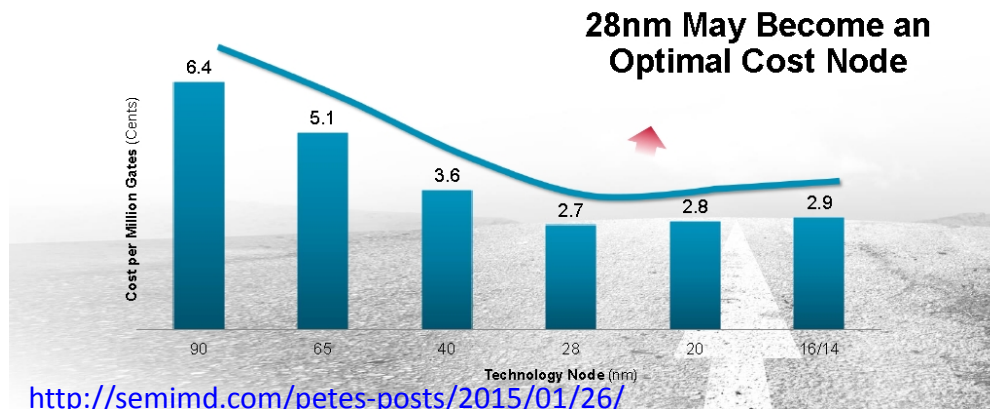
Next: The Bad News



Lennart Johnsson
2015-02-06

Bad News 1: Costs

COST PER TRANSISTOR RISING – HISTORIC FIRST



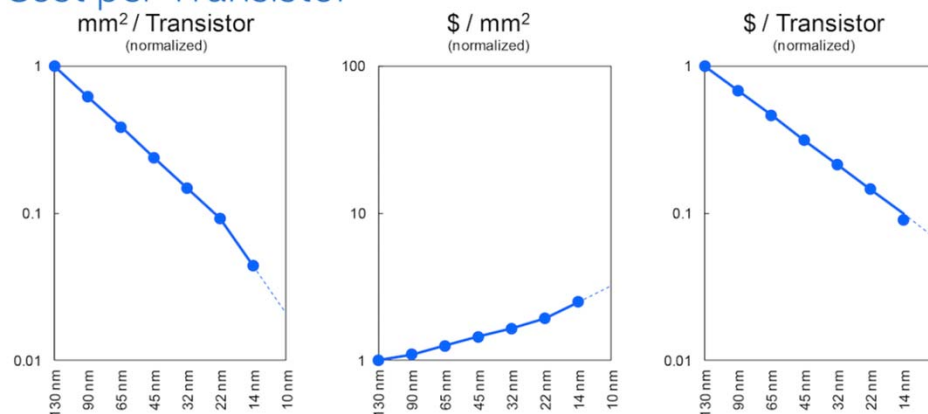
<http://semimd.com/petes-posts/2015/01/26/exponentially-rising-costs-will-bring-changes/>



<http://www.economist.com/news/21589080-golden-be-coming-end-no-moore>

Cost per Transistor

Intel



http://hps.ece.utexas.edu/yale75/mudge_slides.pdf

UNIVERSITY of HOUSTON



Lennart Johnsson
2015-02-06

Bad News 2: Dennard Scaling works no more!

*“Moore’s Law gives us more transistors...
Dennard scaling made them useful.”*



Bob Colwell, DAC 2013, June 4, 2013

Dennard scaling: reducing the critical dimensions while keeping the **electrical field constant** yields **higher speed** and a **reduced power** consumption of a digital MOS circuit

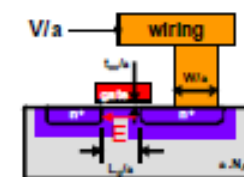
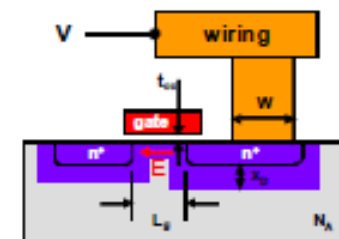
Dennard R.H. et al., “Design of ion-implanted MOSFETs with very small physical dimensions”, IEEE J. Solid-State Circ., vol.9, p.256 (1974)

Using the typical scaling of 0.7x in device dimensions per generation thus results in a doubling of transistors in a given area with the power consumption remaining constant and the speed increasing by 1.4x!!



dimensions t_{ox} , L, W	$1/a$
doping	a
voltage	$1/a$
integration density	a^2
delay	$1/a$
power dissipation/Tr	$1/a^2$
Electric Field E	1

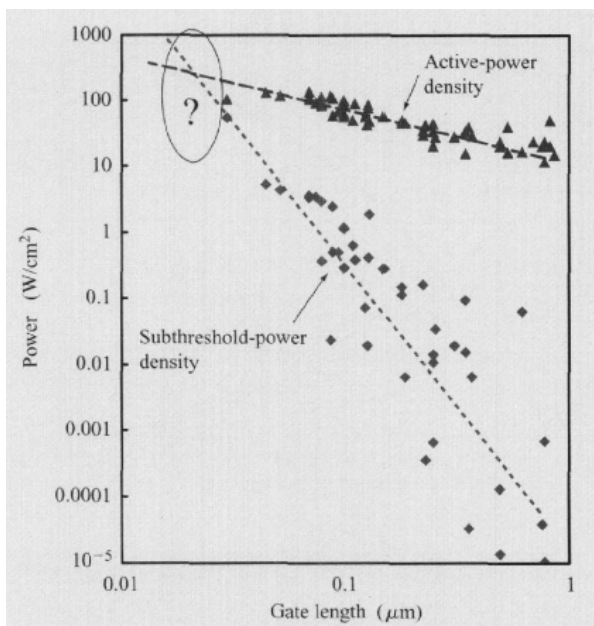
Constant Field Scaling



Dennard scaling ended 2005!

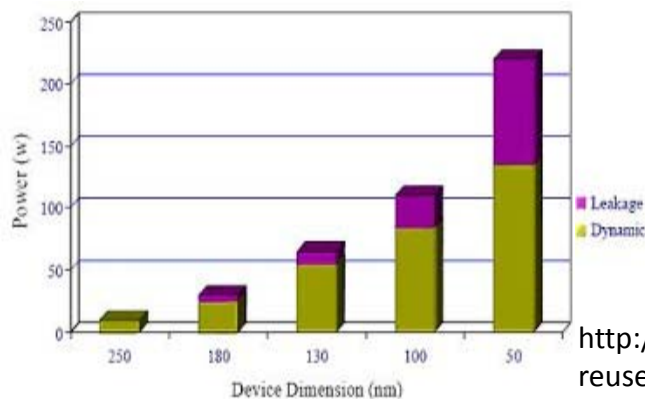
UNIVERSITY of HOUSTON

The end of Dennard Scaling

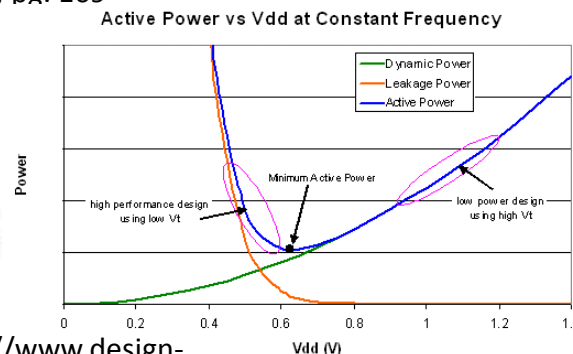


- Dennard scaling did not take subthreshold leakage into account.
- By 2005 subthreshold leakage had increased more than 10,000 fold and further reduction in threshold voltage V_t not feasible limiting operating voltage reduction.
- Further, gate oxide thickness scaling had reached a point of 5 atomic layers and further reduction not possible and direct tunneling current becoming a noticeable part of total chip power.

Graph from E J Nowak IBM Journal of Research and Development, Mar/May 2002, vol. 46, no. 2/3, pg. 169

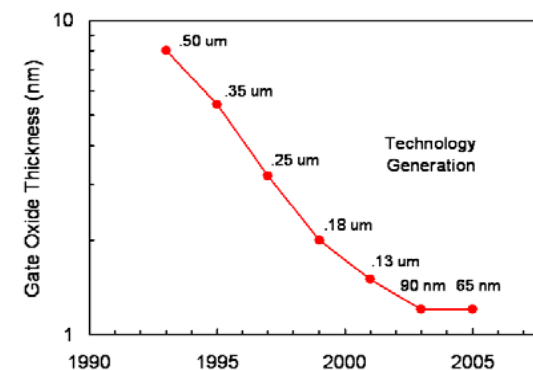


<http://asic-soc.blogspot.com/2008/03/leakage-power-trends.html>



<http://www.design-reuse.com/articles/20296/power-management-leakage-control-process-compensation.html>

UNIVERSITY of HOUSTON

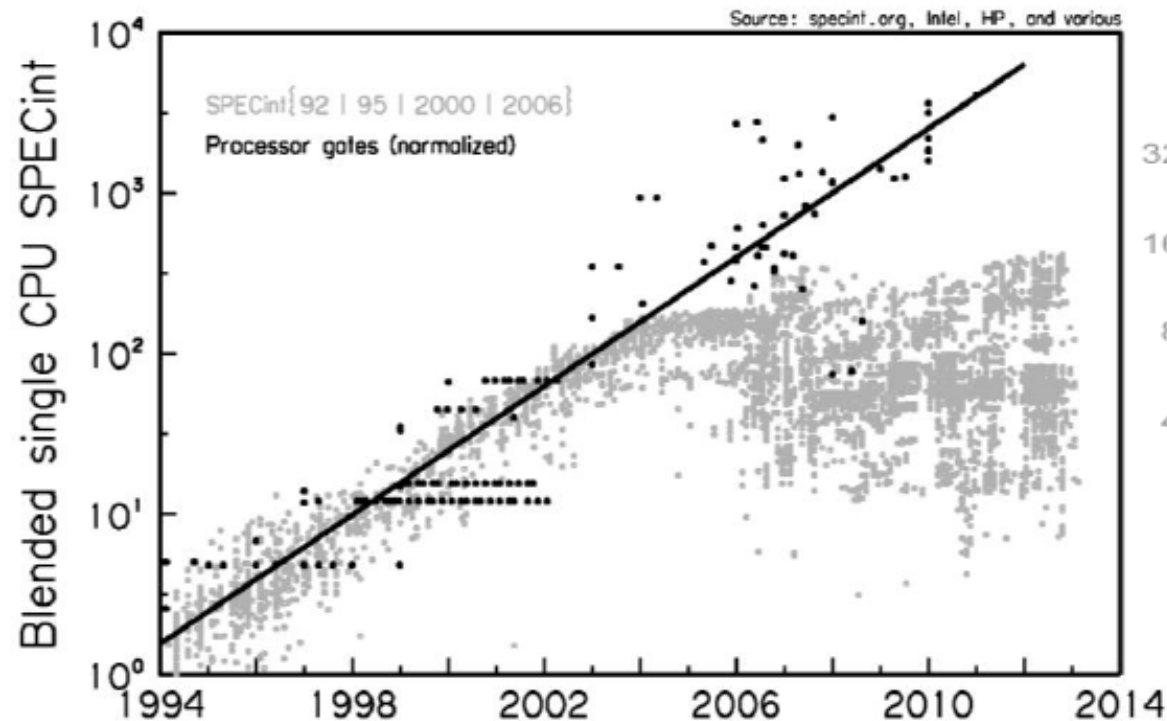


Mark Bohr, A 30-year perspective on Dennard's MPFET Scaling Paper, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4785534

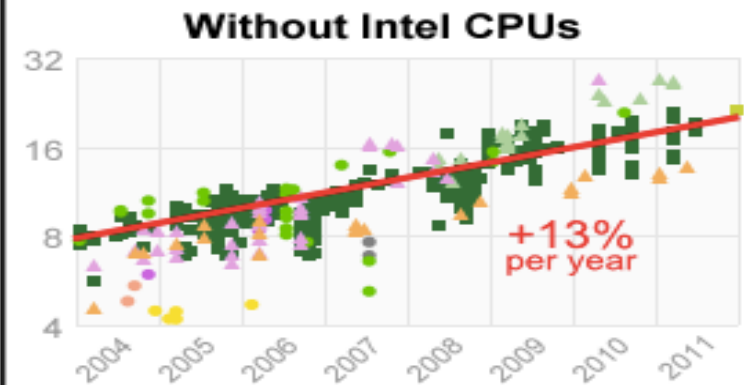


Lennart Johnson
2015-02-06

The End of Dennard Scaling



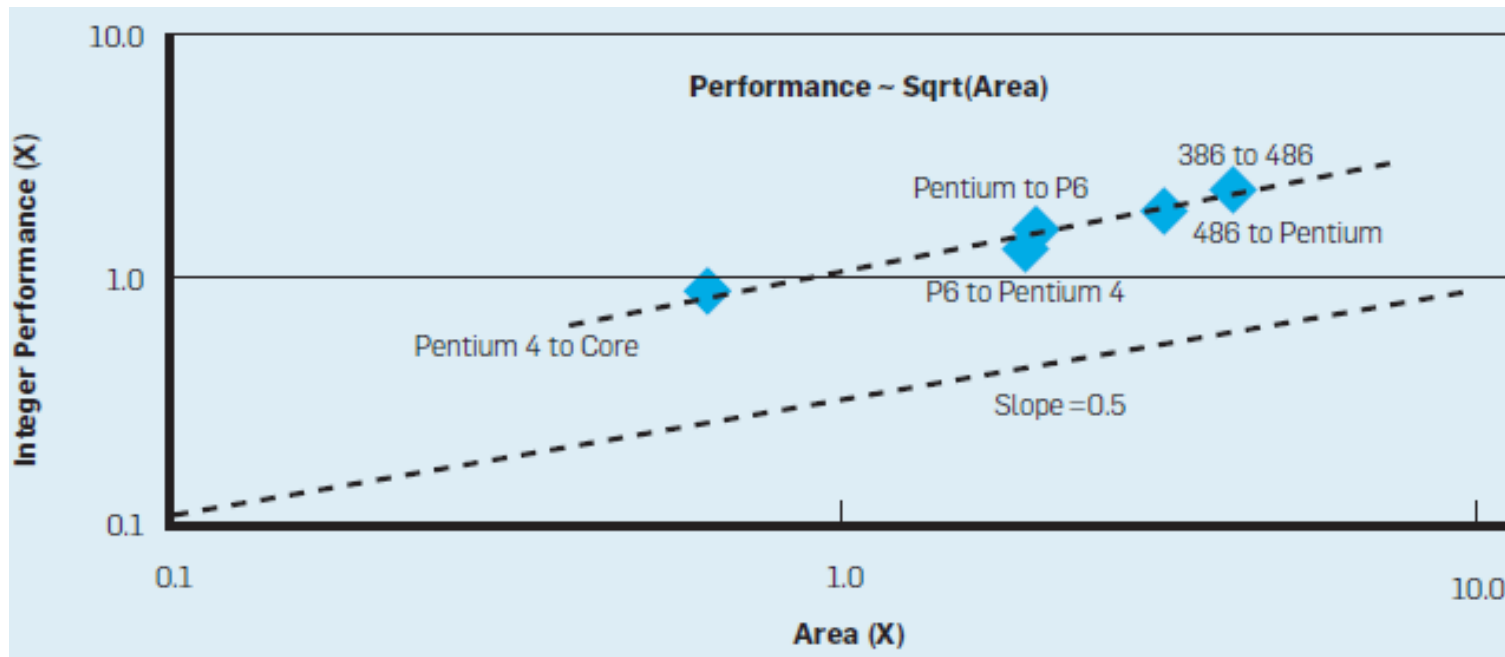
Source: Greg Astfalk <http://www.lanl.gov/conferences/salishan/salishan2013/Astfalk.pdf>



Single thread performance improvement is slow. (Specint)
*”Intel has done a little better over this period, increasing at 21% per year.

Source: Andrew Chien

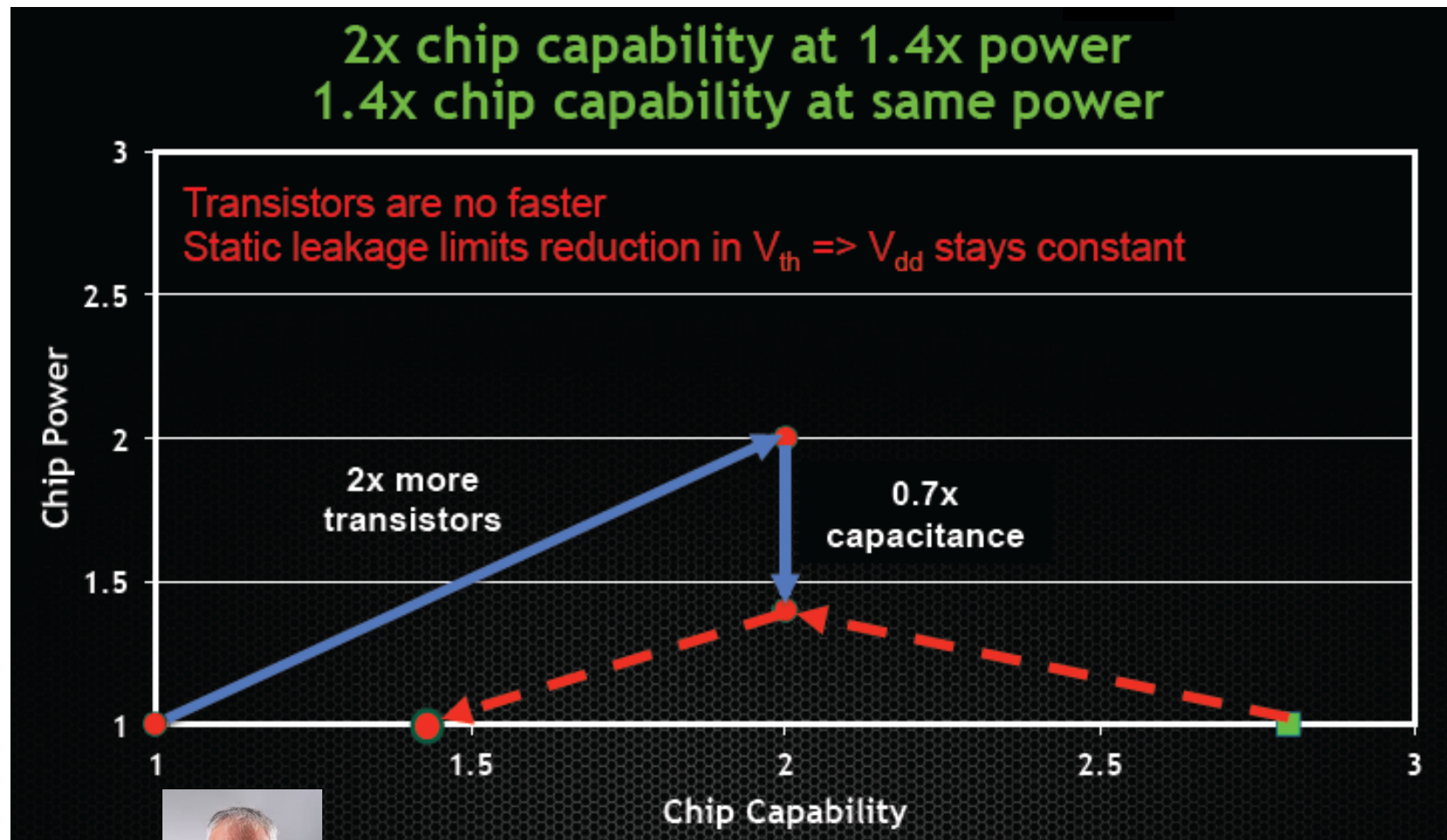
Pollack's rule (Fred Pollack, Intel)



Performance increase @ same frequency is about $\sqrt{\text{area}}$ not proportional to the number of transistors

Source: S. Borkar, A. Chien, The Future of Microprocessors
<http://cacm.acm.org/magazines/2011/5/107702-the-future-of-microprocessors/fulltext>

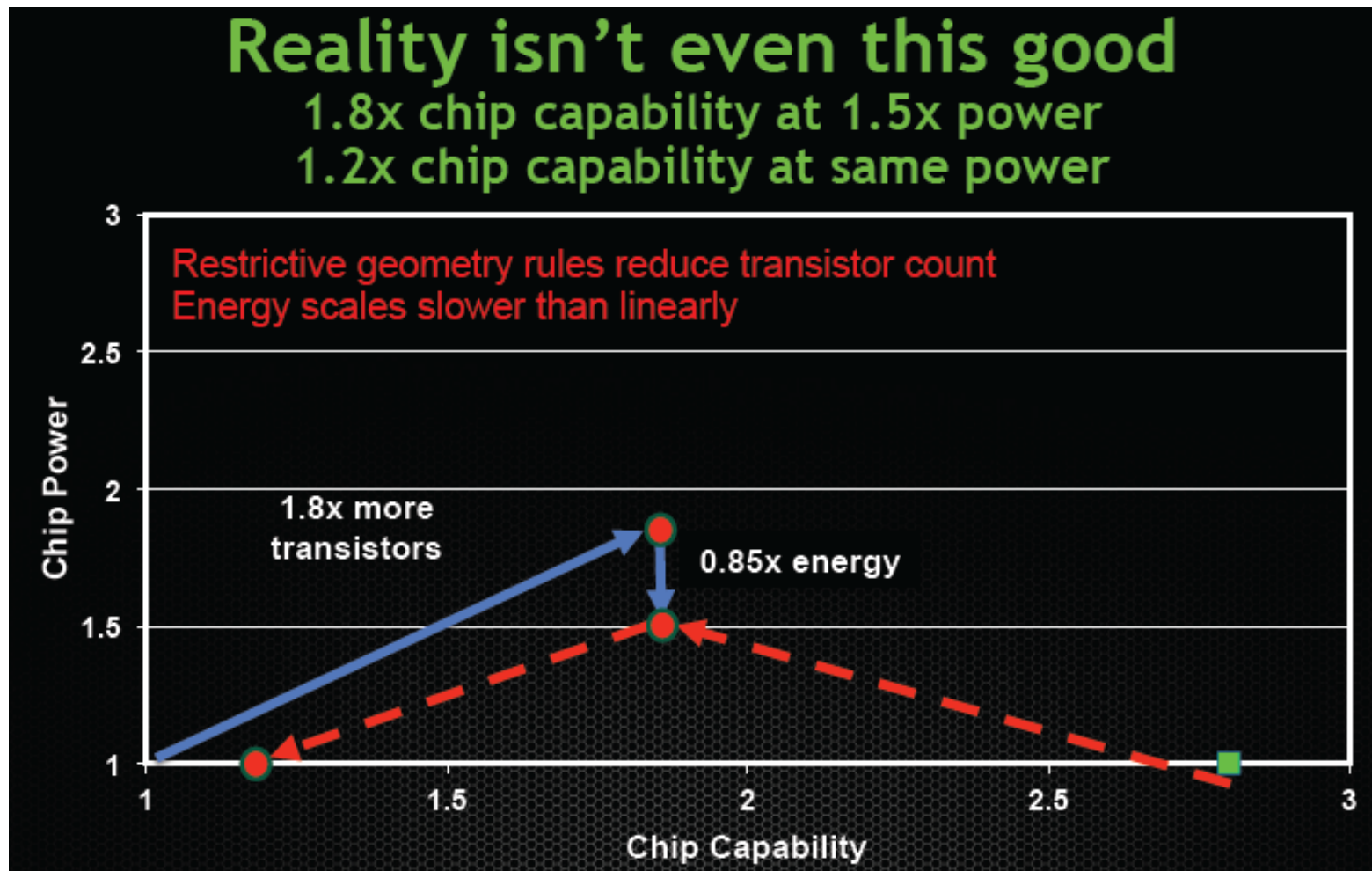
Post Dennard Scaling



Source: Bill Dally, HiPEAC Keynote 2015

UNIVERSITY of HOUSTON

Post Dennard Scaling



Source: Bill Dally, HiPEAC Keynote 2015

UNIVERSITY of HOUSTON



Lennart Johnsson
2015-02-06

The Future

- The conventional path of **scaling** planar CMOS **will face significant challenges set by performance and power consumption** requirements.
- Driven by the 2× increase in transistor count per generation, **power management is now the primary issue** across most application segments. Power management challenges **need to be addressed across multiple levels,** The **implementation challenges of these approaches expands upwards into system design requirements**

2013 Roadmap



International Technology Roadmap for Semiconductors

UNIVERSITY of HOUSTON



Lennart Johnsson
2015-02-06

The Future

“Future systems are energy limited:
Efficiency *is* performance”

“Process matters less: $\sim 1.2x$ per generation”

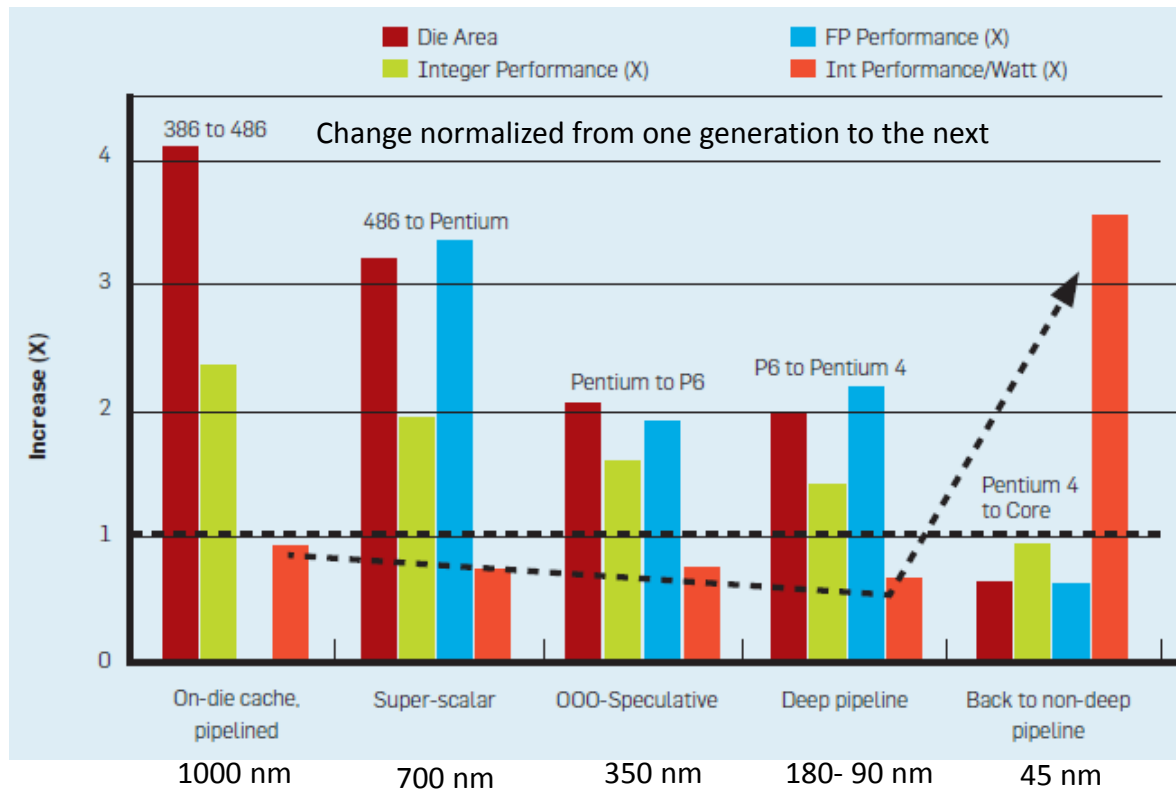
Bill Dally





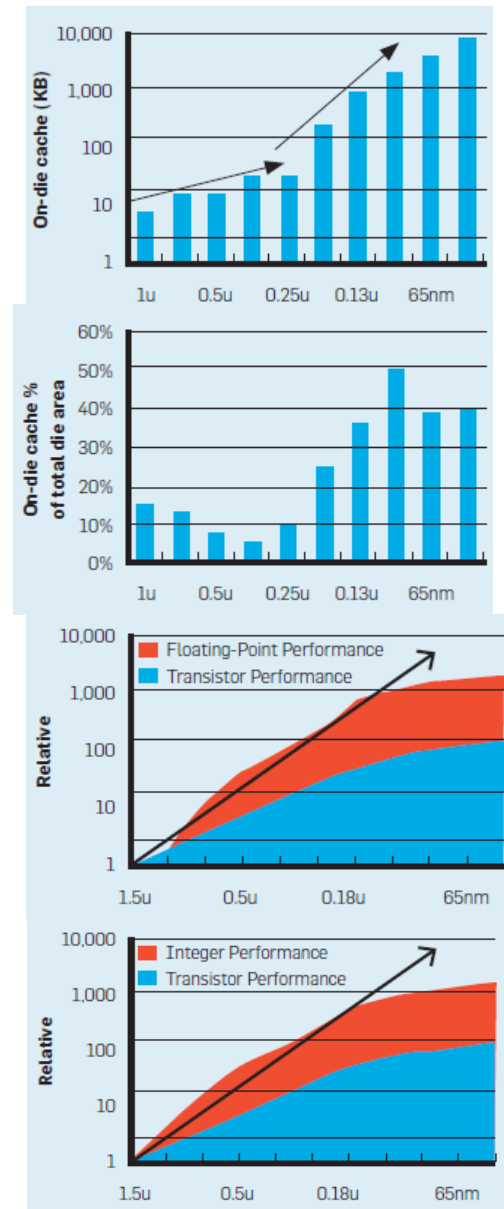
Lennart Johnson
2015-02-06

Retrospective: Microarchitecture benefits



Source: S. Borkar, A. Chien, The Future of Microprocessors

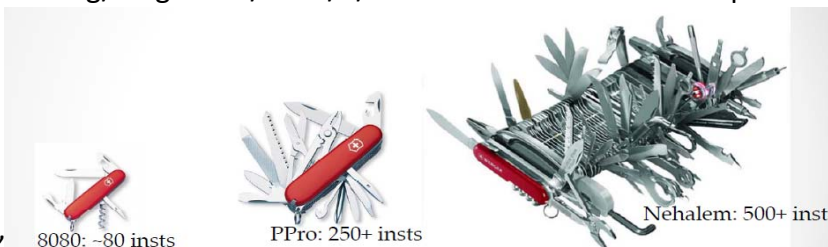
<http://cacm.acm.org/magazines/2011/5/107702-the-future-of-microprocessors/fulltext>



WSOU-ARISTA001491

Source: A. Chien,

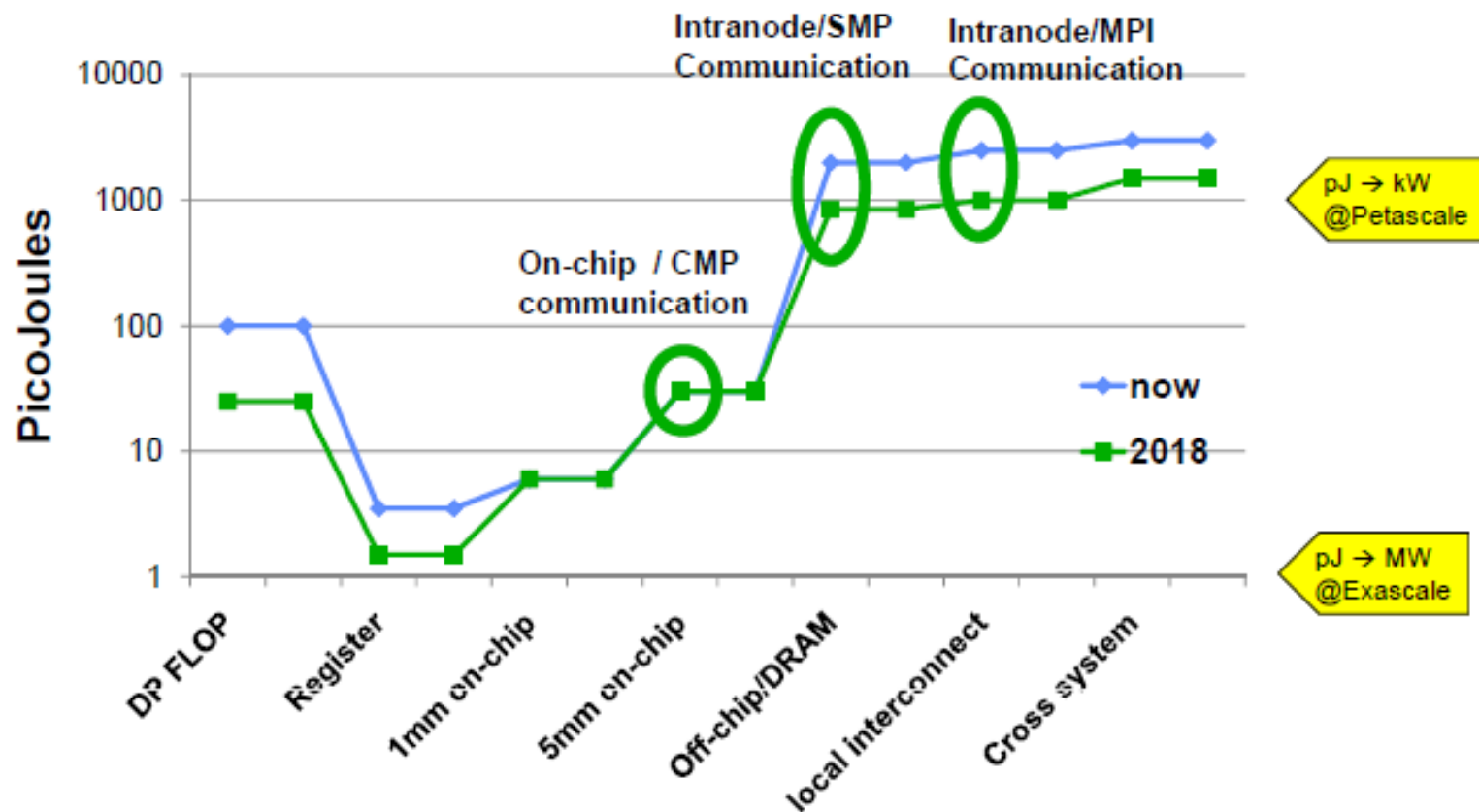
<http://www.lanl.gov/conferences/salishan/salishan2010/pdfs/Andrew%20A.%20Chien.pdf>



UNIVERSITY of HOUSTON



Data Communication



"The Energy and Power Challenge is the most pervasive ... and has its roots in the inability of the [study] group to project any combination of currently mature technologies that will deliver sufficiently powerful systems in any class at the desired levels."
DARPA IPTO exascale technology challenge report

20

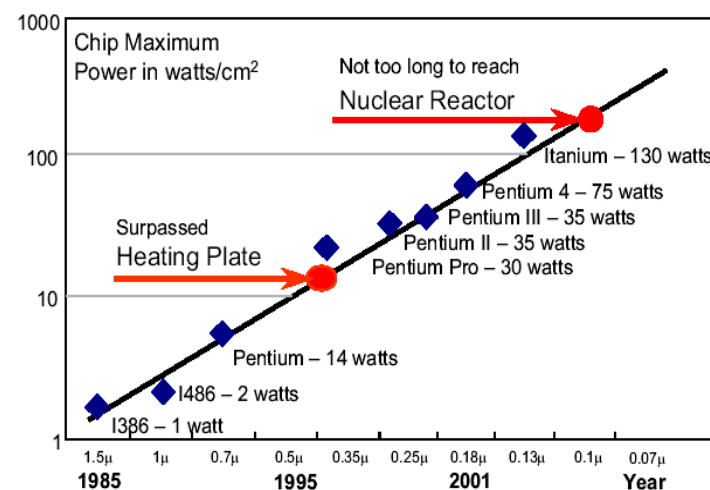


Lennart Johnsson
2015-02-06

Processor Energy History

A retrospective: for several chip generations designers used the increased transistor capability to introduce micro-architectural features enhancing single thread performance, often at the expense of energy efficiency

Product	Normalized Performance	Normalized Power	EPI on 65 nm at 1.33 volts (nJ)
i486	1.0	1.0	10
Pentium	2.0	2.7	14
Pentium Pro	3.6	9	24
Pentium 4 (Willamette)	6.0	23	38
Pentium 4 (Cedarmill)	7.9	38	48
Pentium M (Dothan)	5.4	7	15
Core Duo (Yonah)	7.7	8	11



Source: Shekhar Borkar, Intel

Ed Grochowski, Murali Annavaram Energy per Instruction Trends in Intel® Microprocessors. <http://support.intel.co.jp/pressroom/kits/core2duo/pdf/epi-trends-final2.pdf>

“We are on the Wrong side of a Square Law”

Pollack, F (1999). *New Microarchitecture Challenges in the Coming Generations of CMOS Process Technologies*. Paper presented at the Proceedings of the 32nd Annual IEEE/ACM International Symposium on Microarchitecture, Haifa, Israel.

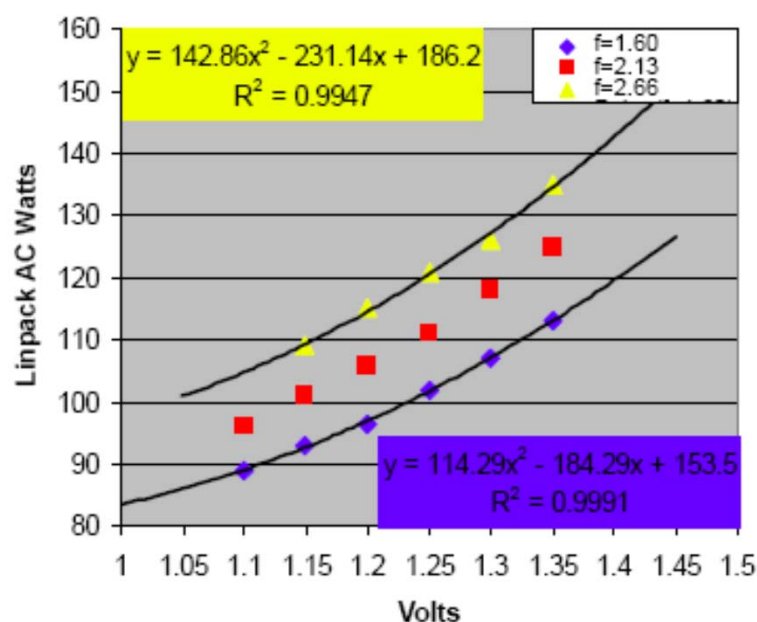


The Square Law

For CMOS the relationship between power (P), voltage (V) and frequency (f) is

$$P = c_1 V^2 f + c_2 V + c_3 + O(V^4)$$

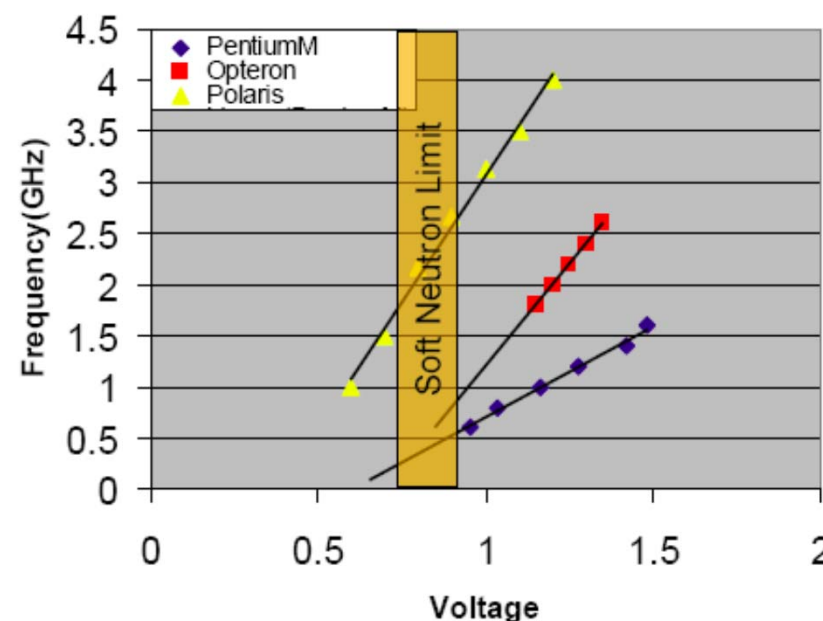
Dynamic Leakage+
Board Fans



$$\text{Linpack: } 15f(V-0.2)^2 + 45V + 19$$

$$\text{STREAM: } 5f(V-0.2)^2 + 50V + 19$$

Furthermore, $f \sim C(V-V_0)$

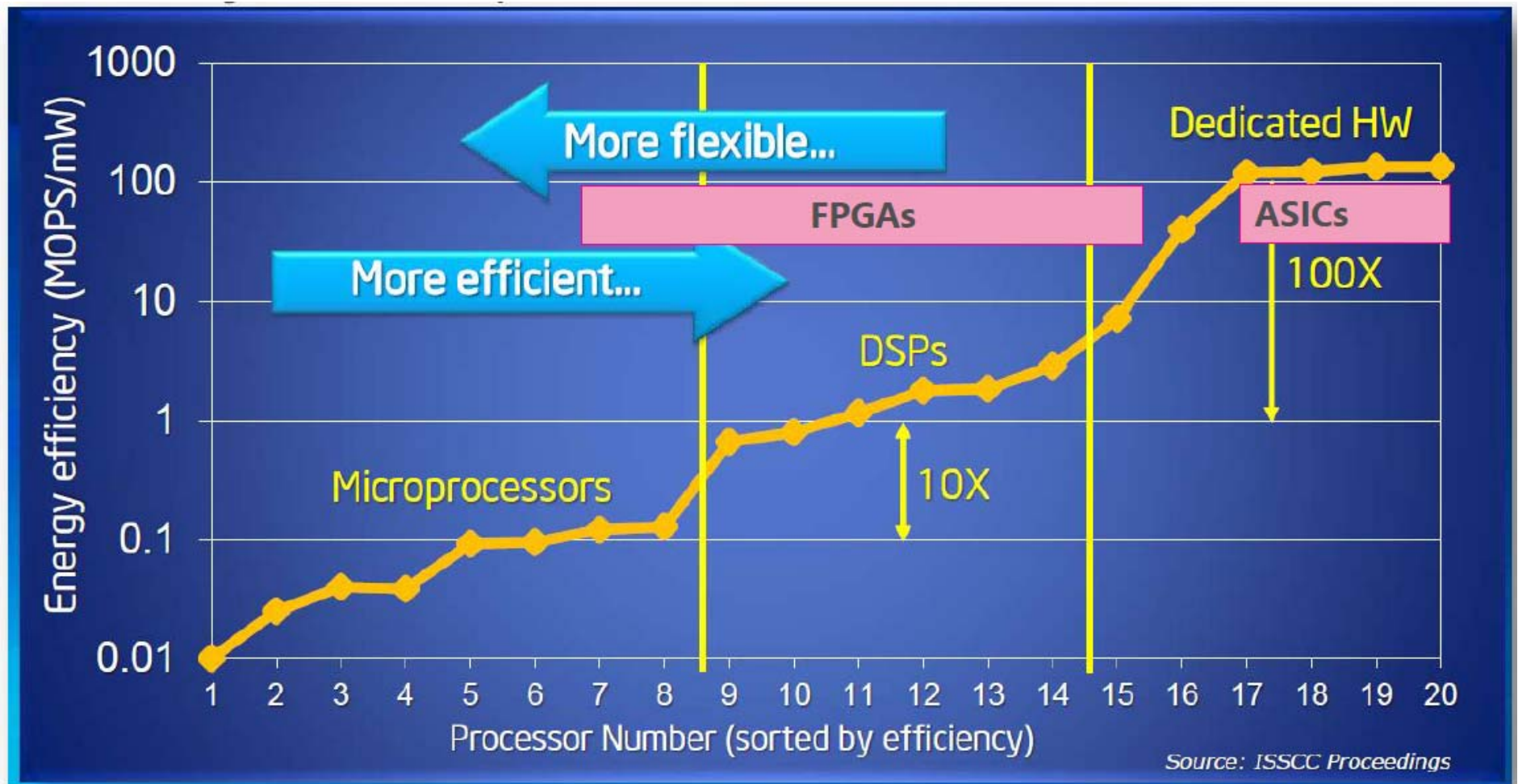


Source: Supermicro



Lennart Johnsson
2015-02-06

Energy Efficiency of different design approaches



http://research.microsoft.com/en-US/events/fs2014/caulfield_adrian_reconfig_fabric_r01.pdf

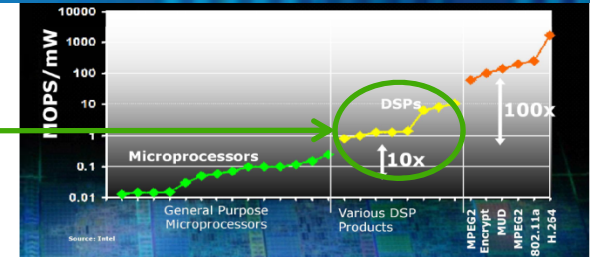
UNIVERSITY of HOUSTON



Lennart Johnson

2015-02-06

Our 2nd Prototype



Source: Andrew Chien,

ARM Cortex-9			ATOM			AMD 12-core			Intel 6-core			ATI 9370		
Cores	W	GF/W	Cores	W	GF/W	Cores	W	GF/W	Cores	W	GF/W	Cores	W	GF/W
4	~2	~0.5	2	2+	~0.5	12	115	~0.9	6	130	~0.6	1600	225	~2.3

nVidia Fermi			TMS320C6678			IBM BQC			ClearSpeed CX700		
Cores	W	GF/W	Cores	W	GF/W	Cores	W	GF/W	Cores	W	GF/W
512	225	~2.2	8	10	~ 4/6	16	55	3.7	192	10	~10

Very approximate estimates!!

KTH/SNIC/PRACE Prototype II



Nominal Compute Density and Energy Efficiency

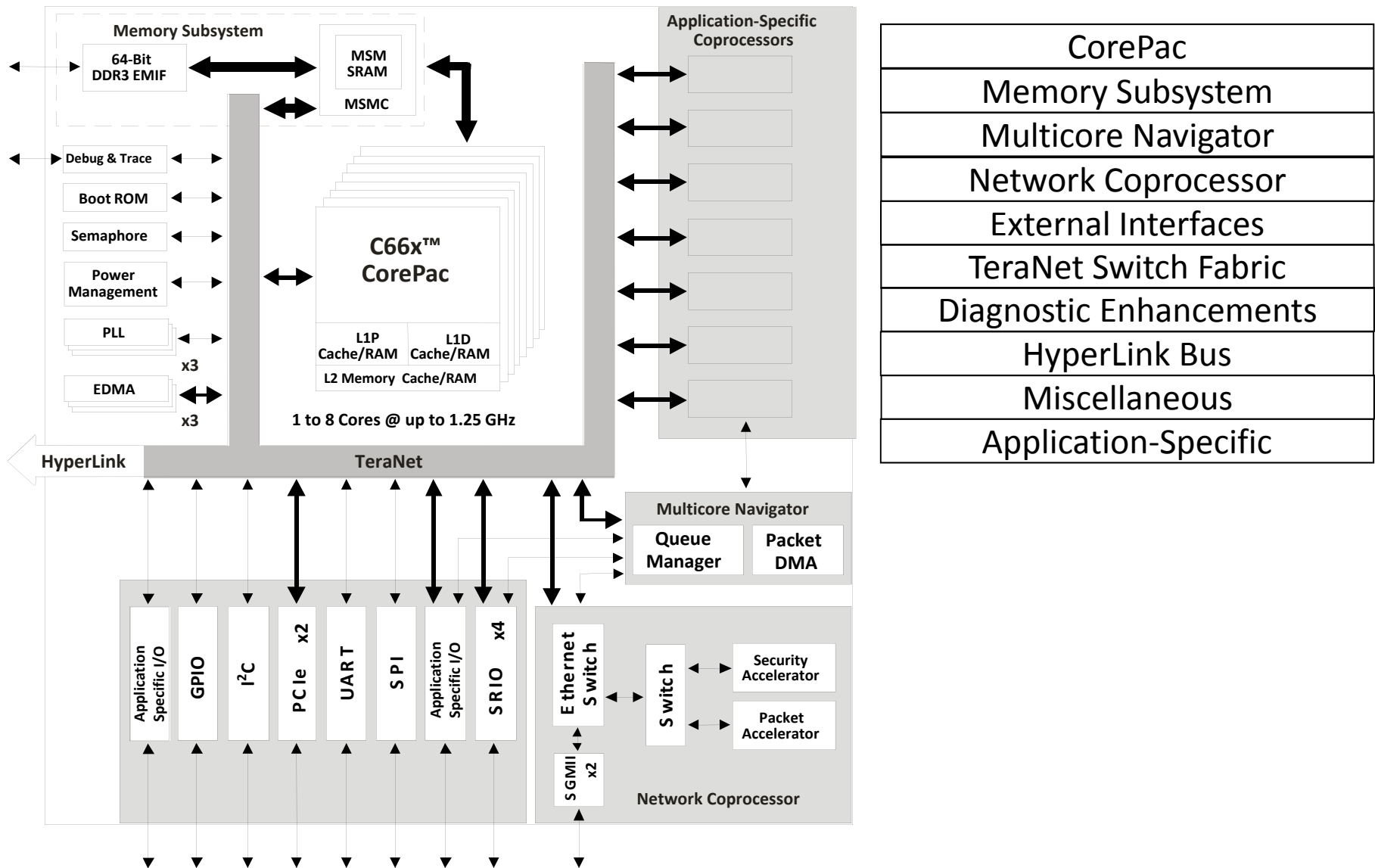
CPU	"Current"			"Next Generation"		
	Feature size	GF/J	GF/mm ²	Feature size	GF/J	GF/mm ²
AMD Interlagos (16C, 2.7 GHz)	32	1.3	0.55			
AMD FirePro 7900	40	3.0	1.19			
AMD S9000				28	3.6	2.21
AMD Brazos	40					
AMD Llano	32					
AMD Trinity	32	1.0	0.40			
IBM Blue Gene/Q	45	3.7	0.57			
IBM Power7	45	1.4	0.48			
Intel Sandy Bridge (8C, 3.1 GHz)	32	1.3	0.46			
Intel Ivy Bridge (4C, 3.5 GHz)				22	1.45	0.70
Nvidia Fermi	40	2.66	1.26			
Nvidia Kepler 20x				28	5.2	2.36
Nvidia Tegra2	40	~1	0.04			
Nvidia Tegra3	40					
TI TMS320C6678	40	4	~3			
TI 66AK2Hx				28		
Xilinx Vertex-6	40	5-10				
Xilinx Vertex-7				28	~13	0.236

Source: L. Johnsson, G. Netzer, Report on Prototype Evaluation, D9.3.3 PRACE, <http://www.prace-ri.eu/IMG/pdf/d9.3.3.pdf>



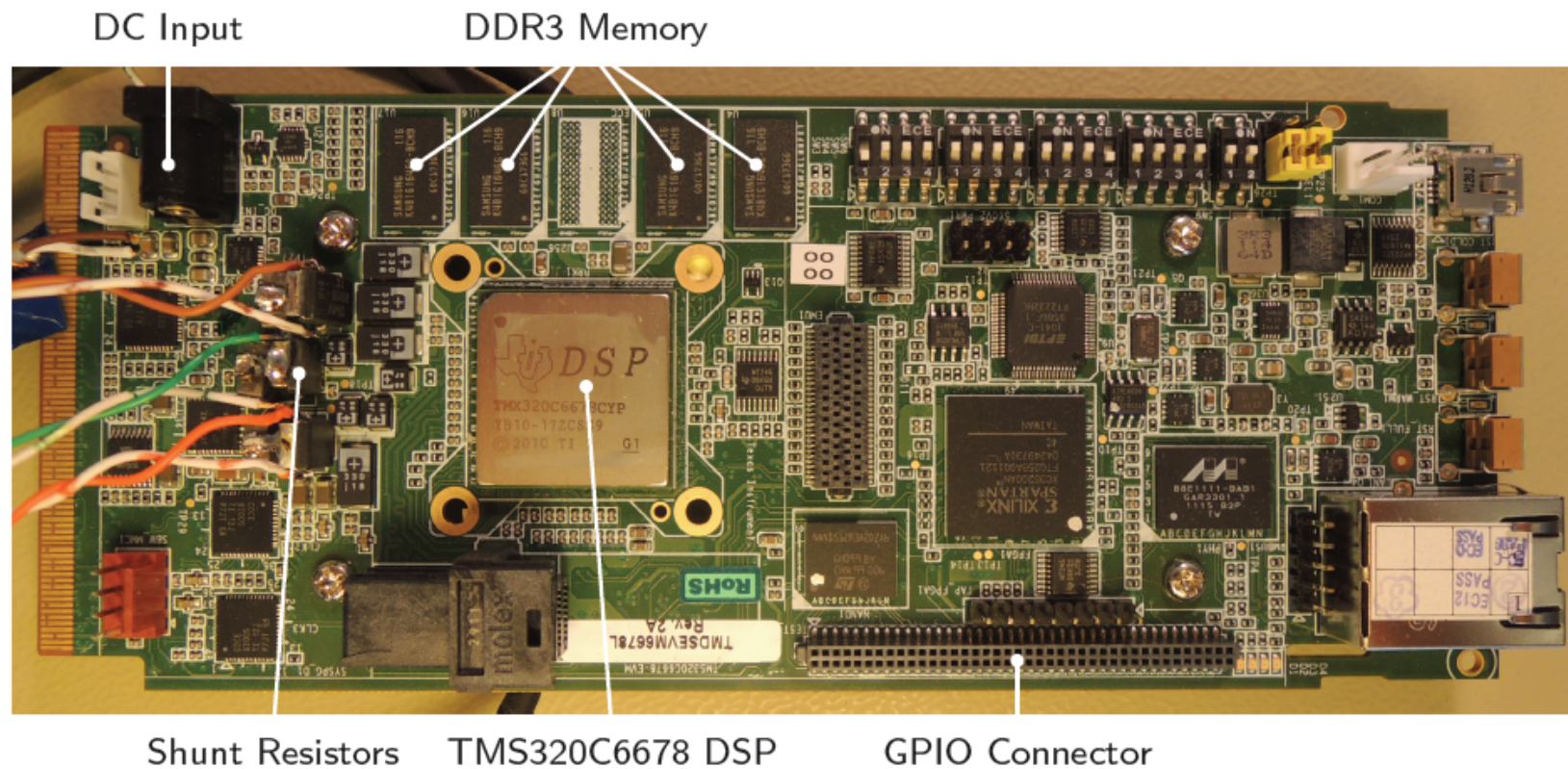
Lennart Johnsson
2015-02-06

TI KeyStone Device Architecture

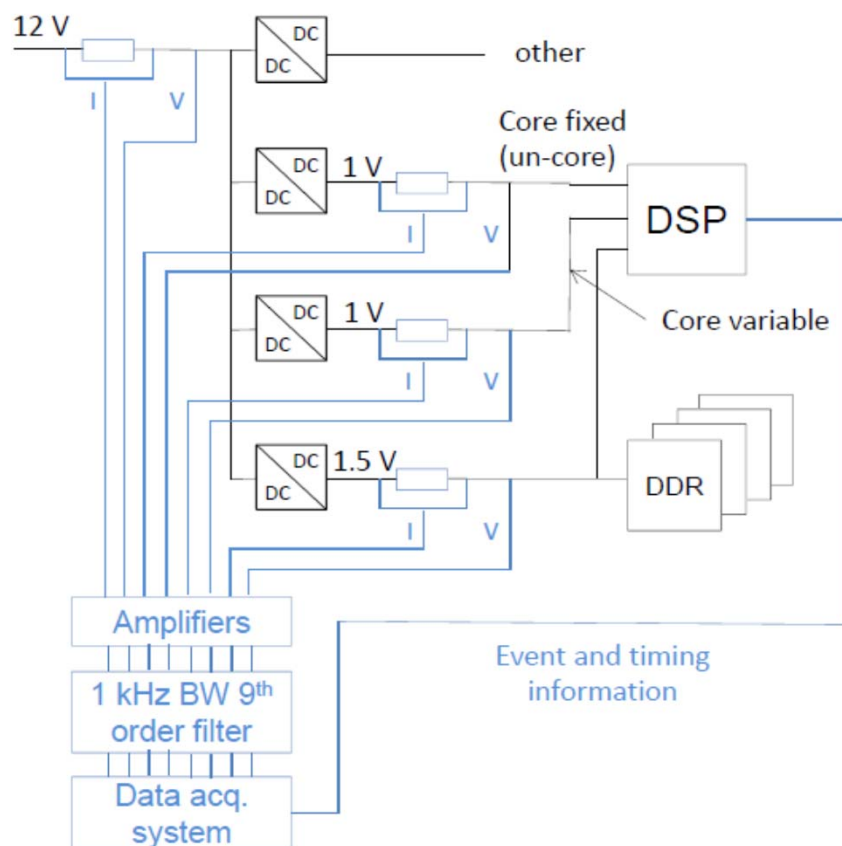


UNIVERSITY of HOUSTON

The TMDXEVM6678L Evaluation Module



- ▶ Added shunt resistors and voltage probes on 4 power feeds.
- ▶ 1 GHz (nom.) 8-core DSP, 1 GiB (512 MiB nom.) DDR3-1333

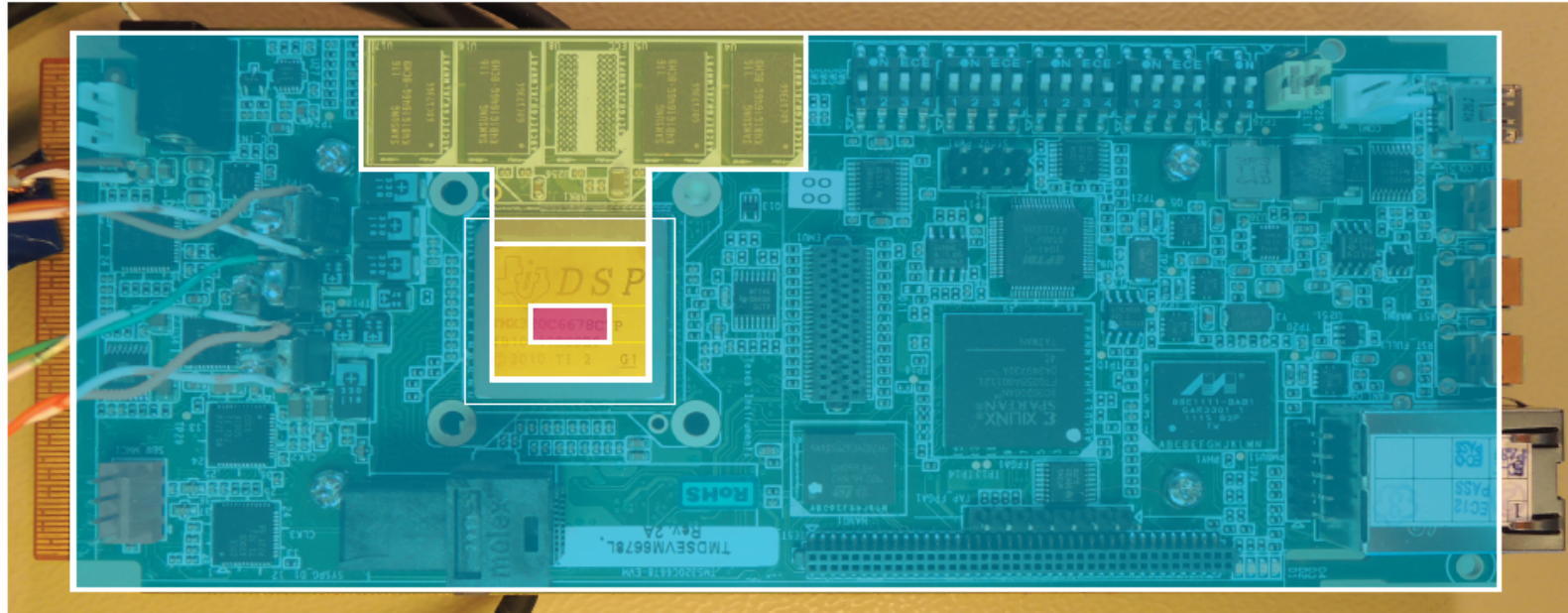






Rail	Voltage	Usage
Variable	0.9 – 1.1 V	All SoC logic
Fixed	1.0 V	On-chip memories Hyperlink Other common fabric
Memory	1.5 V	DDR3 Memory and I/O Hyperlink SerDes PCIe SerDes SGMII SerDes SRIO SerDes
Unit Input	12 V	DC input to EVM (board)
Other		Unit Input – Above Unmeasured parts

Component Error	Relative Error	Contribution
Shunt Resistor ($\Delta T \pm 20K$)	0.1% + 20 ppm/K	1800 ppm
Front End Offset Error	0.01%	2 mV
Front End Gain Error	0.1%	1000 ppm
Front End Passband Flatness	0.1%	1000 ppm
Front End Stopband Rejection	25 ppm	25 ppm
ADC Offset Error	140 ppm	2.4 mV
ADC Gain Error	476 ppm	476 ppm
Total Voltage Error		< 1421 ppm + 9.6 mV
Total Current Error		< 2294 ppm + 19.2 mA
Total Power Error		< 9500 ppm
Marker Jitter (1000s run-time)	4 ppm	4 ppm
Total Energy Error		< 9500 ppm

Measurement Setup

Power Measurement Instrumentation Coverage (1 GHz)



	U_{nom} [V]	2 Cores		8 Cores		Usage
		[W]	%	[W]	%	
 Core Variable	0.9-1.1	3.28	24	5.09	31	SoC logic
 Core Fixed	1.0	0.44	3	0.47	3	SoC memory
 DDR3 Memory	1.5	2.30	17	2.51	15	External memory
 Other	—	7.65	56	8.14	50	Incl. DC-DC losses

Power measurements for 128 MiB STREAM COPY with idle cores power gated.



Results

- STREAM: 96% of peak, 1.47GB/J @ 1GHz
- HPL 77% efficiency, 2.6 GF/W @ 1.25 GHz
 - 95% DGEMM efficiency
- FFT design:
 - 5.2 GF/s @ 1GHz (of 6 GF/s) and 5.6 GB/s for single core for 512 point L1 DP FFT (Bandwidth limited)
 - 20 GF/s @ 1GHz 8 cores for 256k 8L2/MSM DP FFT. Bandwidth limited (max on-chip)
 - 10.3 GF/s @ 1GHz for 128M FFT (DDR3)

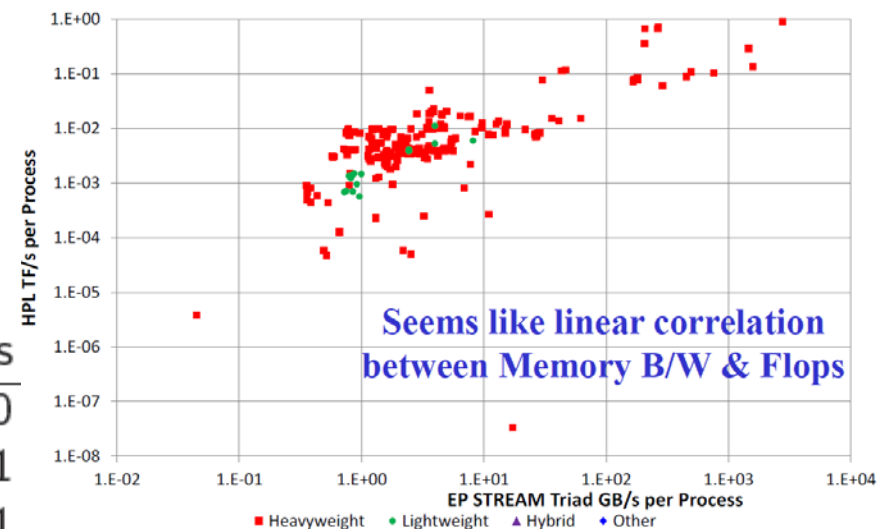


The STREAM Benchmark

- Invented 1990 by John D. McCalpin for assessing (memory) performance of (long) vectors
- Correlates with SPEC-FP performance [McCalpinSC02]
- Operations

Name	Operation	Data (s_i) [B]	FLOPs
COPY	$a_i = b_i$	16	0
SCALE	$a_i = qb_i$	16	1
SUM	$a_i = b_i + c_i$	24	1
TRIAD	$a_i = b_i + qc_i$	24	2

How Does HPL Relate to Streams?



Source: Peter M Kogge, IPDPS 2014



6678 STREAM results in perspective

Platform	BW (GB/s)		Eff. %	Energy Eff. (GB/J)	Lith. nm	Comment
	Peak	Meas.				
IBM Power7	409.6	122.8	30.0	≈0.10	45	STREAM power est. at 50% of TDP
Intel Xeon Phi	352.0	174.8	49.7	≈1.20	22	STREAM power est. at 50% of TDP
Intel E5-2697v2	119.4	101.5	85.0	≈0.30	22	STREAM power est. at 50% of TDP
NVIDIA Tegra3	6.0	1.6	26.7	0.05	40	Power and BW measured (BSC)
6678 Cache	10.7	3.0	28.1	0.40	40	Power and BW measured at 1 GHz
6678 EDMA	10.7	10.2	95.7	1.26	40	Power and BW measured at 1 GHz

IBM Power7 data from “IBM Power 750 and 760 Technical Overview and Introduction”, Literature Number 4985, IBM, May 2013 and W.J.Starke, “Power7: IBMs next generation, balanced POWER server chip,” HotChips 21, 2009.

Intel Xeon Phi data from R. Krishnaiyer, E. Kultursay, P. Chawla, S. Preis, A. Zvezdin, and H. Saito, “Compiler-based Data Prefetching and Streaming Non-temporal Store Generation for the Intel R Xeon Phi Coprocessor”, 2013 IEEE 27th International Parallel and Distributed Processing Symposium Workshops&PhD Forum, NewYork, NY, IEEE, May 2013, pp. 1575–1586.

Intel E5-2697v2 data from Performance and Power Efficiency of Dell Power Edge servers with E5-2600v2, Dell, October, 2013



STREAM Challenges and Solutions

- STREAM operations have no data reuse.
 - Worst case for caches, only merging accesses into cache blocks.
 - Write-allocate policies increase memory traffic.
 - Evicted data shares address LSBs with new data leading to DRAM memory bank conflicts.
- Possible to implement fast STREAM using:
 - Prefetching to hide long latencies.
 - Bypassing cache on writes to prevent extra traffic.
- Make the data transport explicit in the program:
 - Can use block-copy engines.
 - Can schedule accesses.
 - Can utilize regularity in access patterns.



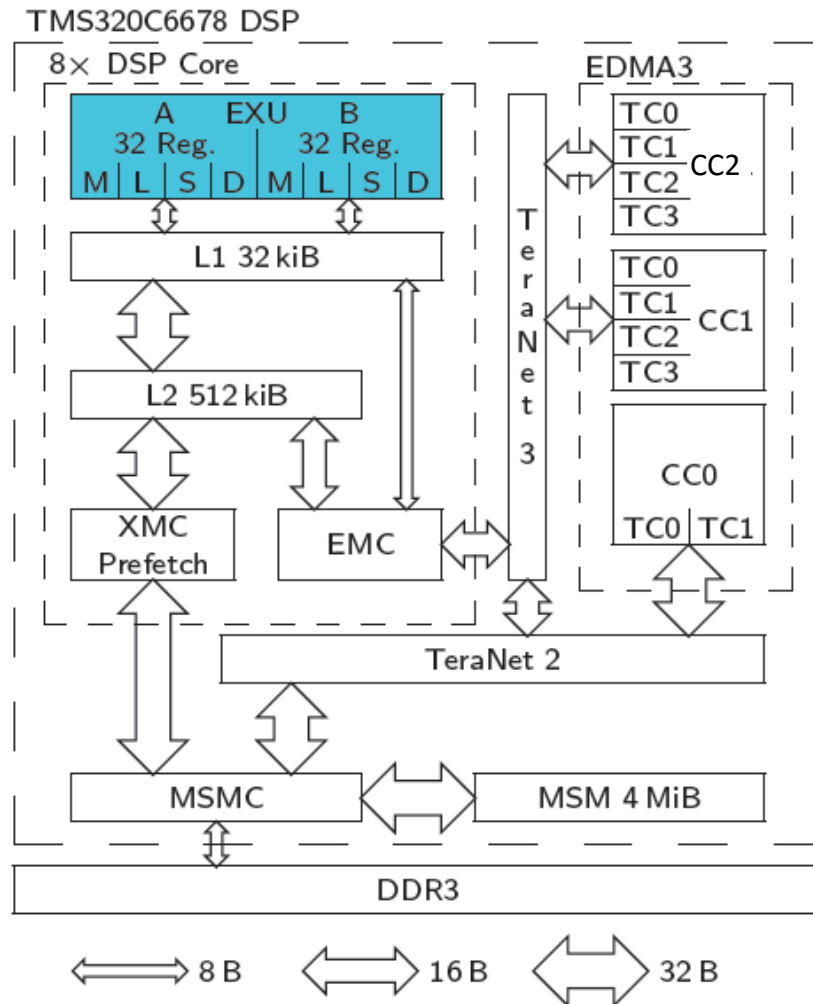
Lennart Johnsson
2015-02-06

Understanding and Optimizing STREAM on the 6678



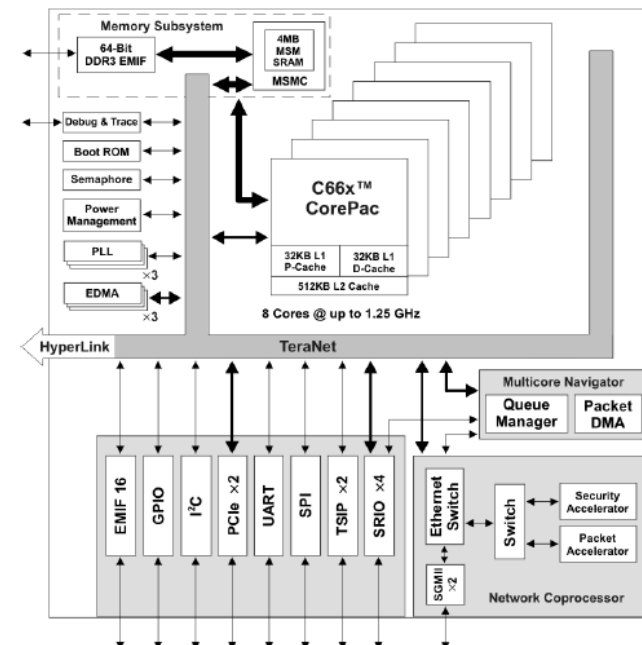
Lennart Johnsson
2015-02-06

The Texas Instruments TMS320C6678 DSP



The TMS320C6678 multi-core DSP:

- ▶ 8 cores (0.8-1.25 GHz) with:
 - ▶ 8-way VLIW **Execution Unit**, 2 64-bit load or stores, 4 FLOPs per cycle



UNIVERSITY of HOUSTON

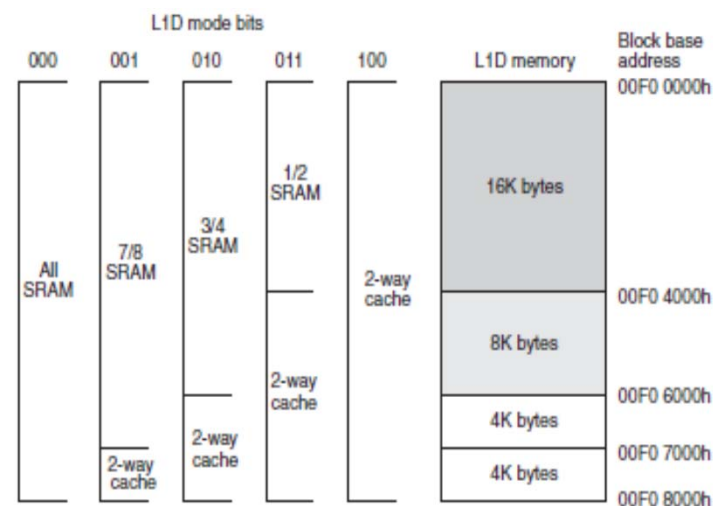
WSOU-ARISTA001507



Lennart Johnsson
2015-02-06

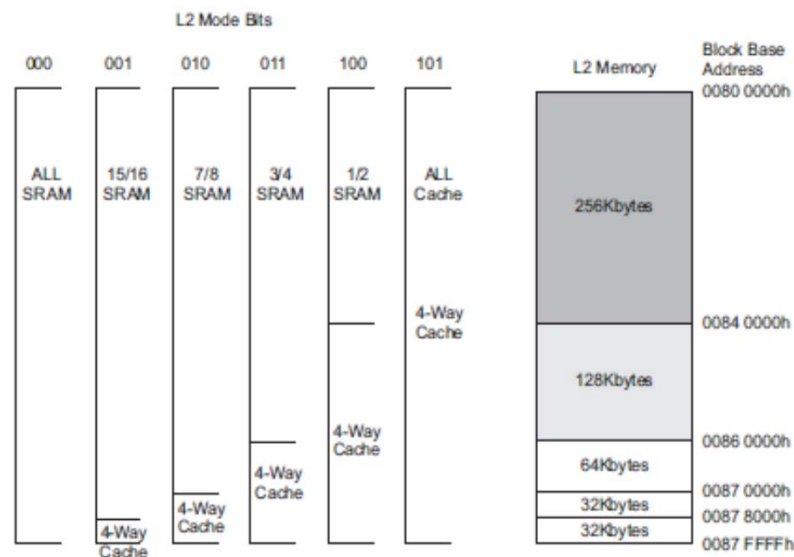
The 6678 Core Memory

	Unit	L1I	L1D	L2
Total Size	KiB	32	32	512
Cache Block Size	B	32	64	128
Associativity		Direct	2-way	4-way
Replacement Policy			LRU	LRU
Allocation Policy		Read	Read	Read/Write
Clock		f_{CK}	f_{CK}	$f_{CK}/2$
Latency SRAM	Cycles	6	5	8–15.5
Latency Cache	Cycles	6	5	12–17.5
Bandwidth	B/Cycle		16	16



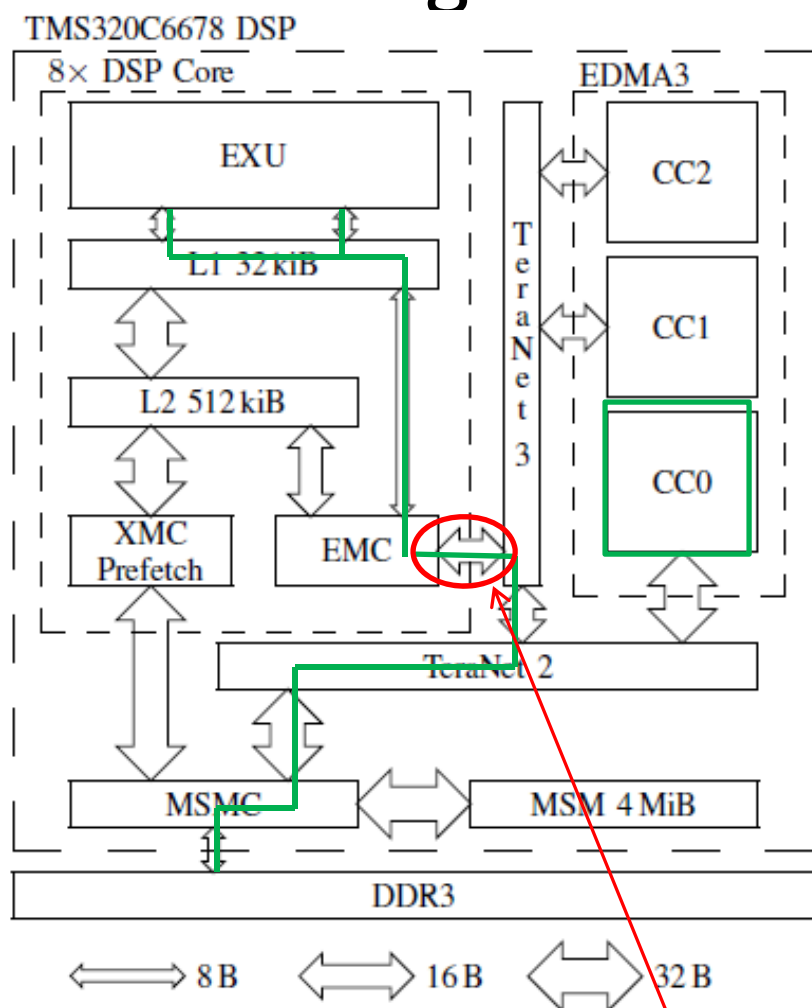
- The 6678 has a 16B 4-entry write buffer for bypassing the L1 on cache misses to avoid stalls, if the write buffer is not full. The buffer drain rate is:

- 16B at CPU/2 rate to L2 SRAM
- 16B at CPU/6 rate to L2 cache





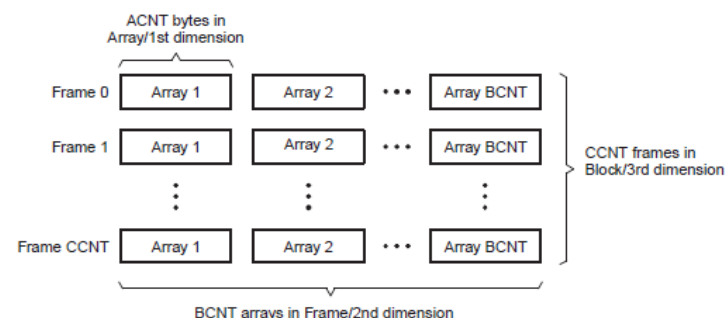
Avoiding Stalls and Hiding Latency using the Enhanced DMA (EDMA3)



Bandwidth limiting path 16/3 B/CPU cycle

EDMA3

Three Dimensional Transfers



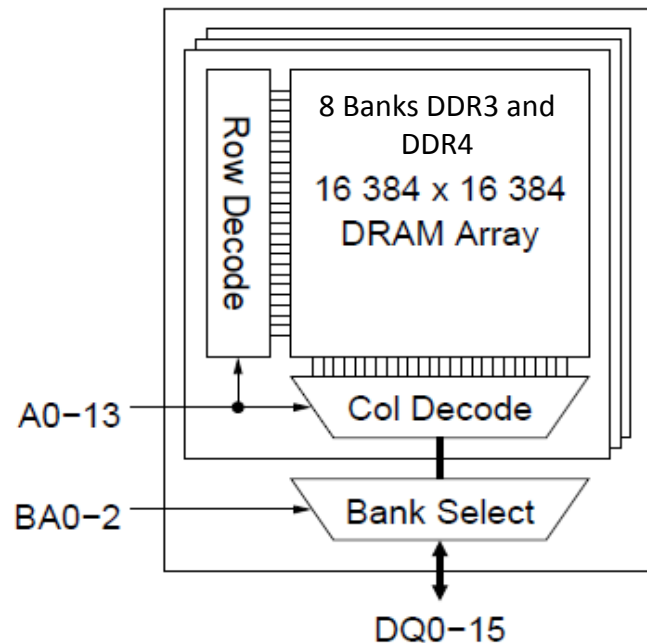
- Three Channel Controllers (CCx), with CC0 having 2 Transfer Controllers (TCs), and CC1 and CC2 each having 4 TCs
- Support for *chaining*, completion of a transfer triggers a subsequent transfer, and *linking*, automatic loading of transfer specifications
- Uses TeraNet 2, CPU clock/2, and TeraNet 3, CPU clock/3

- CC0 used in our EDMA3 STREAM impl.h

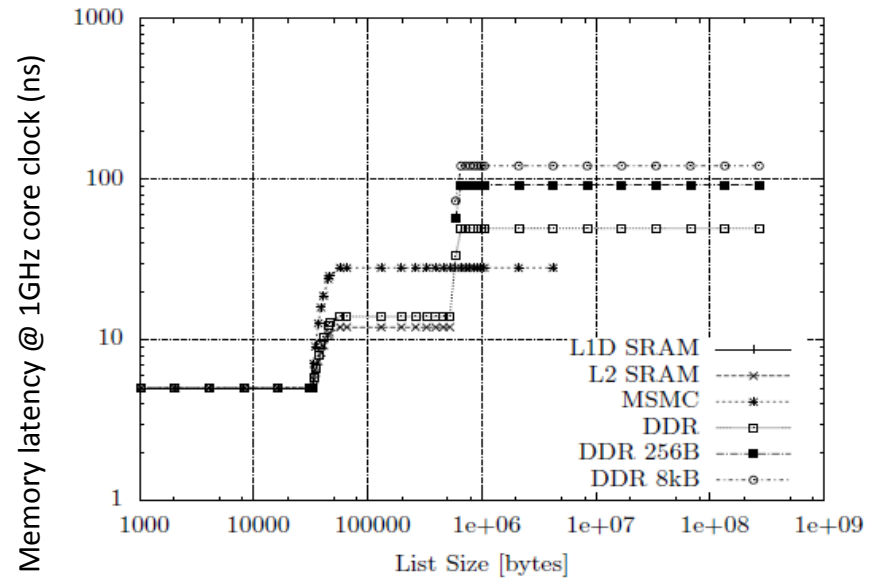


Lennart Johnsson
2015-02-06

The DDR Challenge: Avoid Page Faults



DDR3-1333	t_{burst} (ns)	BW (GB/s)
Page	6	10.66
4 Banks	11.25	5.68
1 Bank	49.5	1.29



Memory	Stride in Bytes								
	64	128	256	512	1 ki	2ki	4ki	8ki	16ki
L1 SRAM	5	5	5	5	5	5	5	5	5
L2 SRAM	12	12	12	12	12	12	12	12	12
MSM	15	28	28	28	28	28	28	28	28
DDR3-1333	28	49	92	93	95	99	106	121	122

This is a serious problem for multi-core CPUs!

See e.g. *Collective Memory Transfers for Multi-Core Chips*,

G. Michelogiannakis, A. Williams, and J. Shalf, Lawrence Berkeley National Lab, November 2013.

Measured latencies at 1 GHz core clock.

L2 cache line: 128 B. Prefetch distance: 256 B

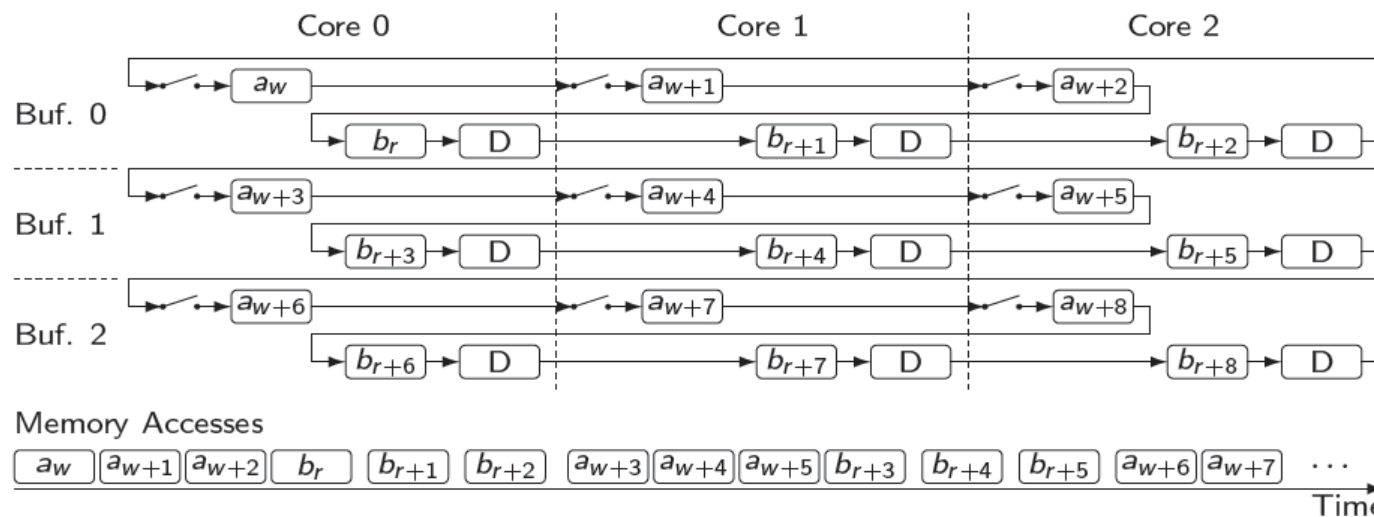
DDR3 page size 8kiB.



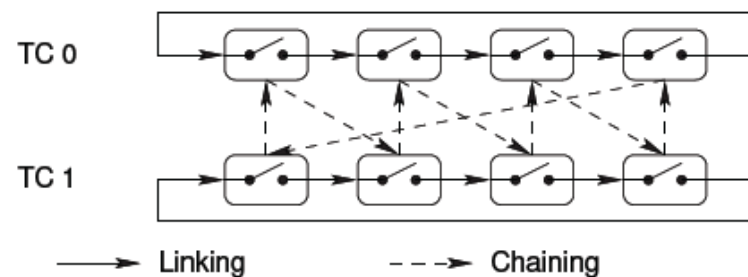
EDMA3 Strategy

- Prefetch to L1 directly from DDR3
 - Eliminates stalls and hides latency
- Store from L1 directly to DDR3
 - Eliminates stalls, hide latencies, and avoids write-allocate reads
- Use at least two buffers per core to enable overlap of core Execution Unit operations with data transfers (due to various overheads more than two buffers may be beneficial). (If the core execution time exceeds the transfer time, then the core-EDMA3 interface will not be fully utilized even with multiple buffers.)
- Use the EDMA3 to make core operations entirely local (no inter-core communication required for scheduling, synchronization, etc.)
- Use the EDMA3 to coordinate core accesses to DDR3 to maximize bandwidth utilization by minimizing DDR3 page openings and closings
- Order DDR3 reads and writes to minimize read/write and write/read transitions
- Use two Transfer Controllers to maximize memory channel utilization (overlap TC overheads)

Scheduling EDMA3 transfers



Clustering of loads and stores across cores reduces read/write and write/read transitions

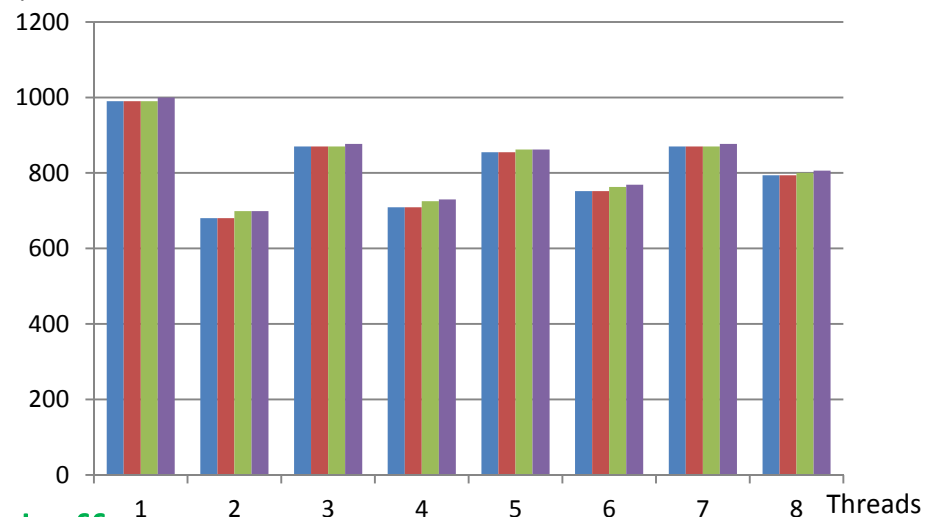
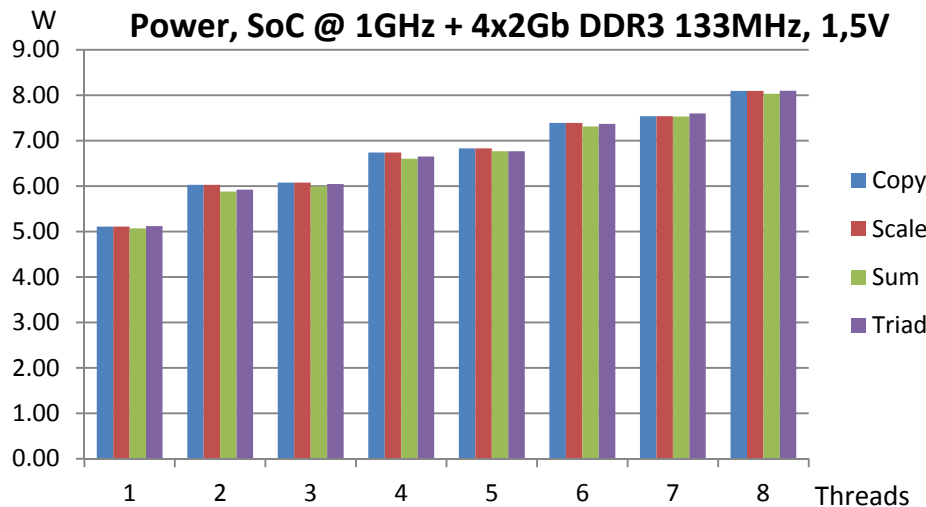
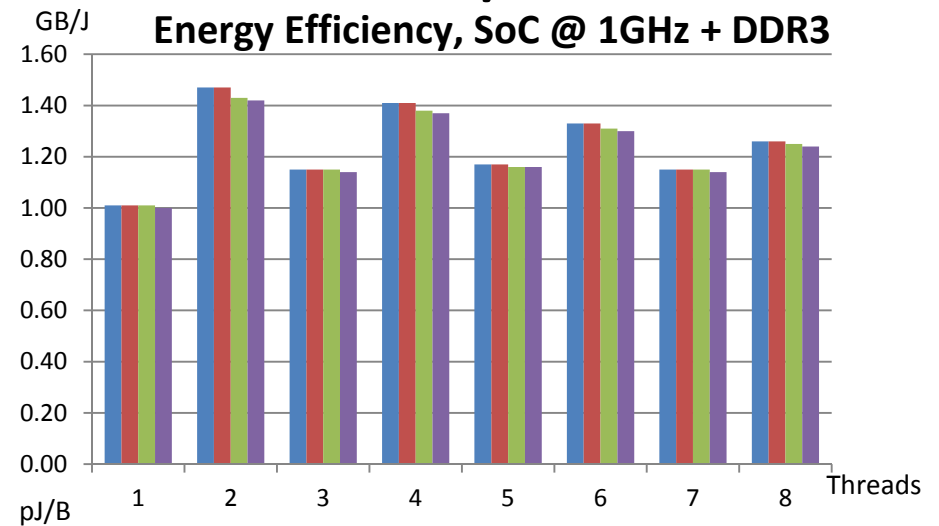
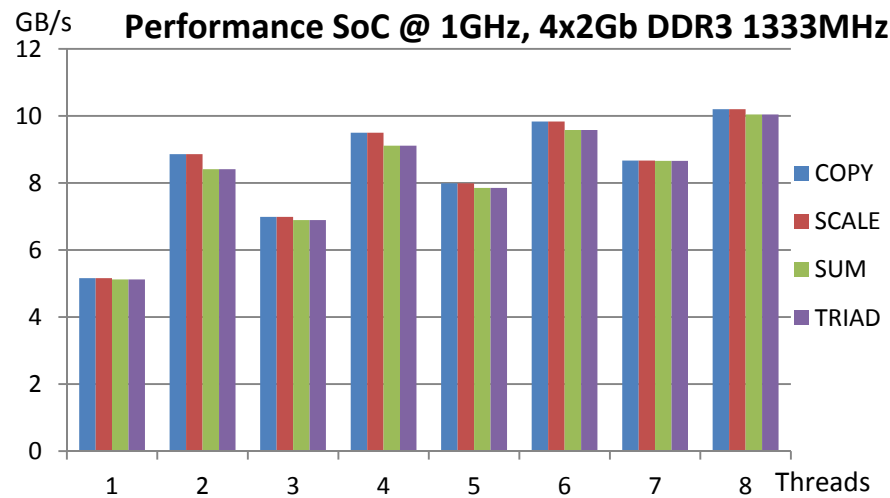


Independent Transfer Controllers are kept in lock-step using cross-triggering



Lennart Johnson
2015-02-06

STREAM Power and Energy Efficiency, EDMA3, 2 TCs, 3 or 2 buffers/core



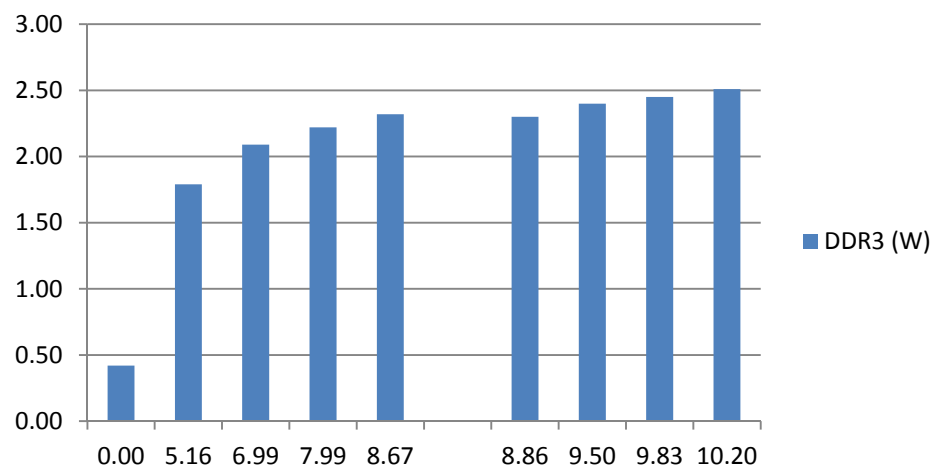
Idle cores powered off

UNIVERSITY of HOUSTON

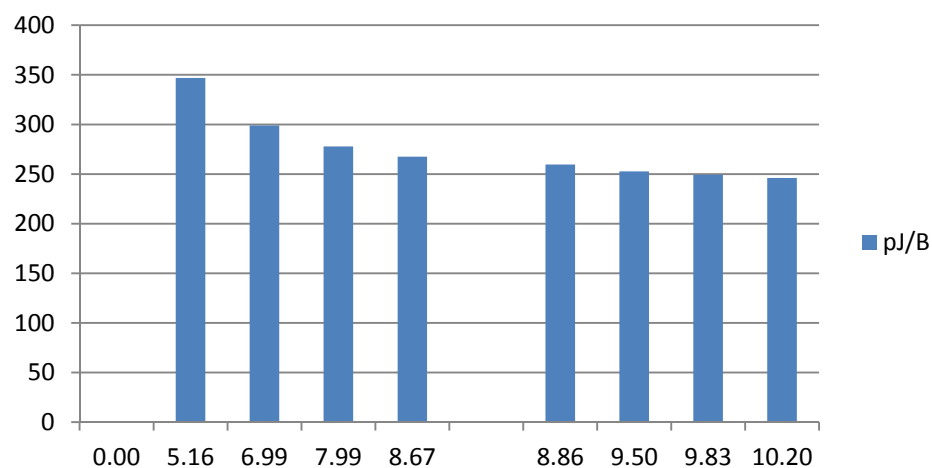


DDR3 Memory Power and Energy

W, 4 x 2Gb DDR3-1333, 1.5V



pJ/B, 4 x 2Gb DDR3-1333, 1.5V





Lessons Learned from Optimizing STREAM

- The 40 nm 6678 DSP can offer energy efficiency comparable to 22 nm x86 CPUs with complex hardware prefetch and hardware support for non-temporal stores (bypassing write-allocate mechanisms).
- Explicit data management necessary (using the EDMA3) to implement
 - effective prefetching
 - storing to DDR3
 - coordinate core accesses to DDR3
 - multi-buffering to overlap core and transfer operations
 - overlap Transfer Controller overheads by using multiple (two) TCs
- The EDMA3 linking and chaining features needed to achieve close to 100% memory channel utilization
- Optimization at the level carried out for our STREAM implementation requires detailed and deep knowledge of the architecture, its features and limitations, but for highly structured applications encapsulating this complexity into higher level functions or libraries seems feasible (but was not attempted within the time available).



Lennart Johnsson
2015-02-06

The Future

Key Limits for Chip Design

1985-2005

- Transistor Count
- Design Complexity
- Valid.&Test Complexity
- Clock Rate
- ~~Energy~~

2005-2020?

- ~~Transistor Count~~
- Design Complexity
- Valid.&Test Complexity
- ~~Clock Rate~~
- Energy

2020+

- ~~Transistor Count~~
- ~~Design Complexity~~
- ~~Valid.&Test Complexity~~
- ~~Clock Rate~~
- Energy

Andrew Chien, VP of Research, Intel

<http://www.lanl.gov/conferences/salishan/salishan2010/pdfs/Andrew%20A.%20Chien.pdf>



Dark silicon

- Chip power will remain capped at current levels (because of cooling costs and known cooling technologies)
- Moore's law enables more transistors per unit area, but post Dennard scaling power per unit area increases
- Thus, **dark silicon**, i.e. not all areas can be simultaneously used, **becomes a necessity**
- **On-die dynamic power management** (voltage, frequency, clock, power) will be a necessity. For maximum performance and energy efficiency this may propagate up to the application level.

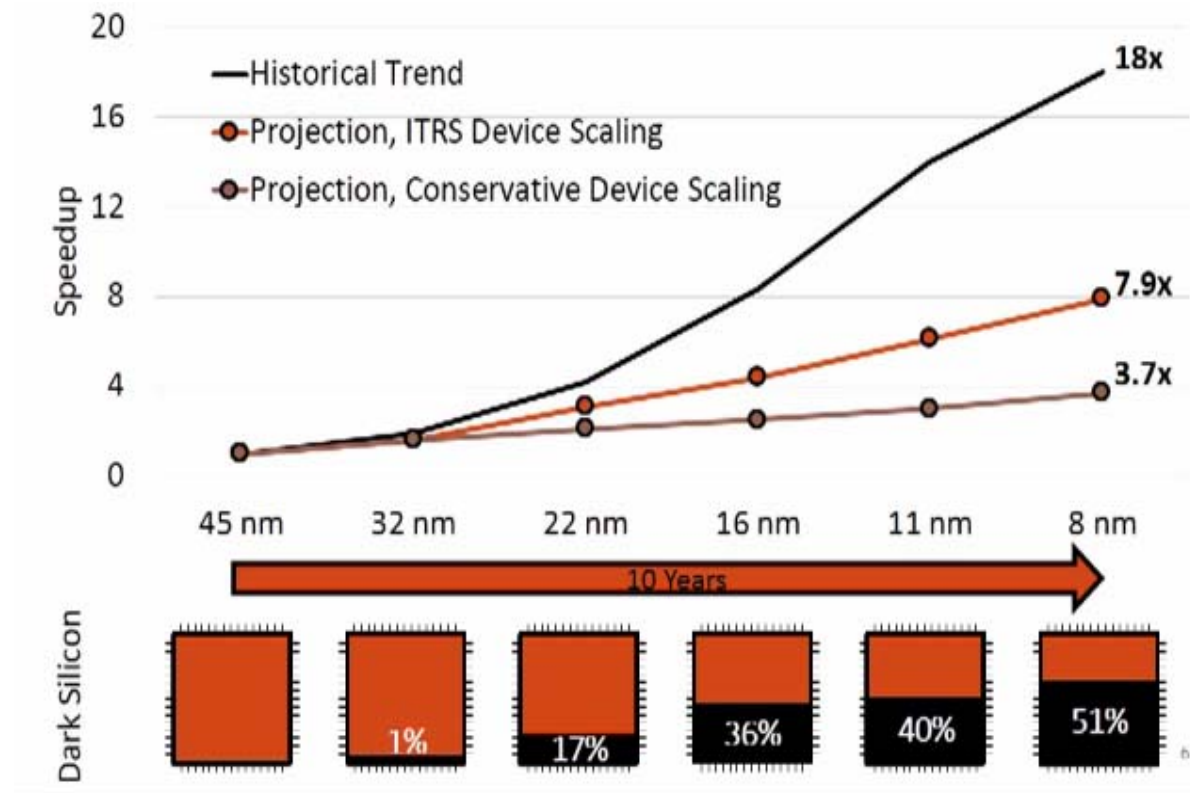
Remark 1. New technology has made it possible for voltage control to move onto the die, enabling increasingly refined control of subdomains of a die

Remark 2. Recent processor chips has a dedicated processor for power control, including control of voltage, frequency, power and clock distribution to subdomains of the chip.



Lennart Johnsson
2015-02-06

Power Challenges



Power Challenges May End the Multicore Era Hadi Esmaeilzadeh, Emily Blem, Renée St. Amant, Karthikeyan Sankaralingam, Doug Burger,
<http://dl.acm.org/citation.cfm?id=2408797&CFID=745900626&CFTOKEN=86138313>

<https://www.youtube.com/watch?v=Df8SQ8ojEAQ&feature=youtu.be>

UNIVERSITY of HOUSTON

“The good news is that the old designs are really inefficient, leaving lots of room for innovation,”



Bill Dally, Nvidia/Stanford,
NYT, July 31, 2011



UNIVERSITY of HOUSTON



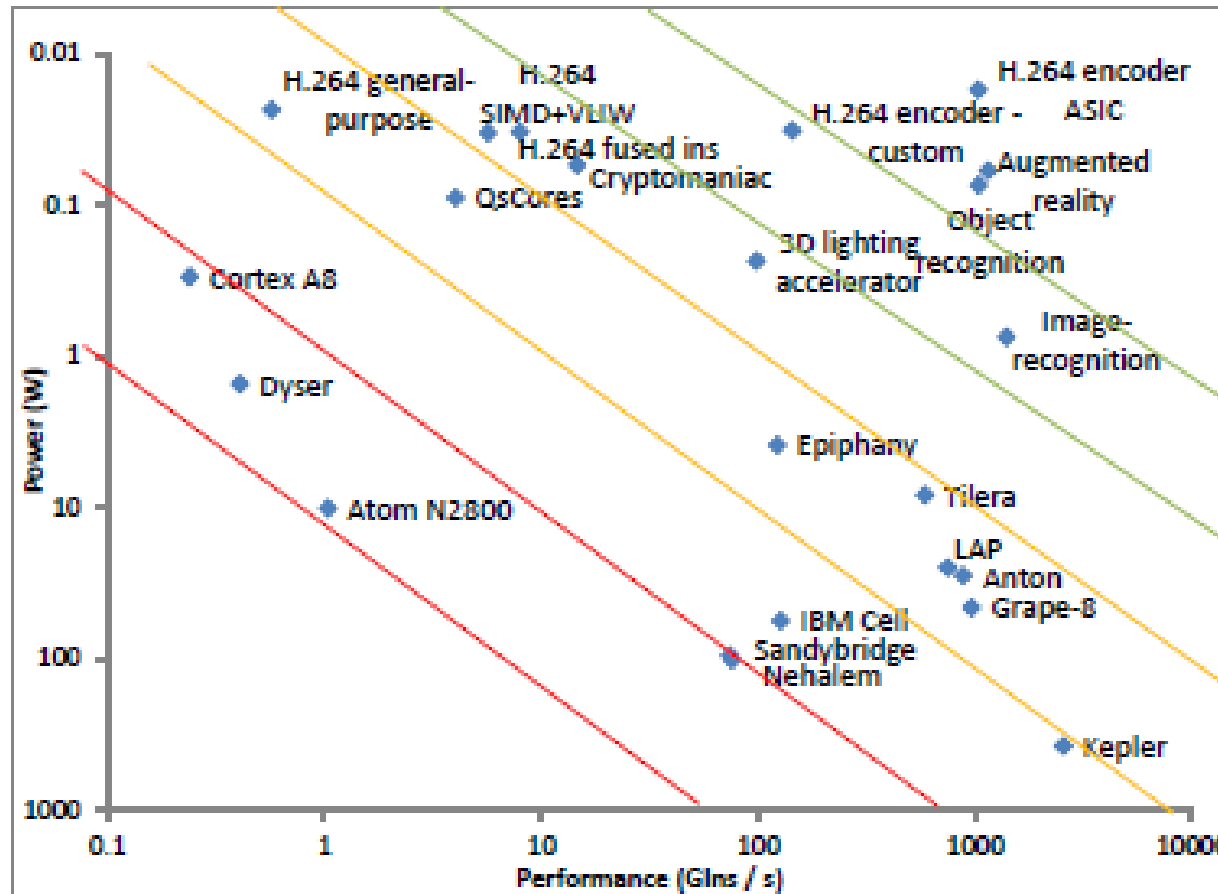
Lennart Johnsson
2015-02-06

“We used to focus graduate students on generating ideas of how to improve performance by adding features, now we need to focus them on how to improve performance by removing features.”

Bill Dally,
A Computer Architecture Workshop:
Visions for the Future, September 19, 2014

Customization Power/Performance Examples

Note: lowest power consumption on top



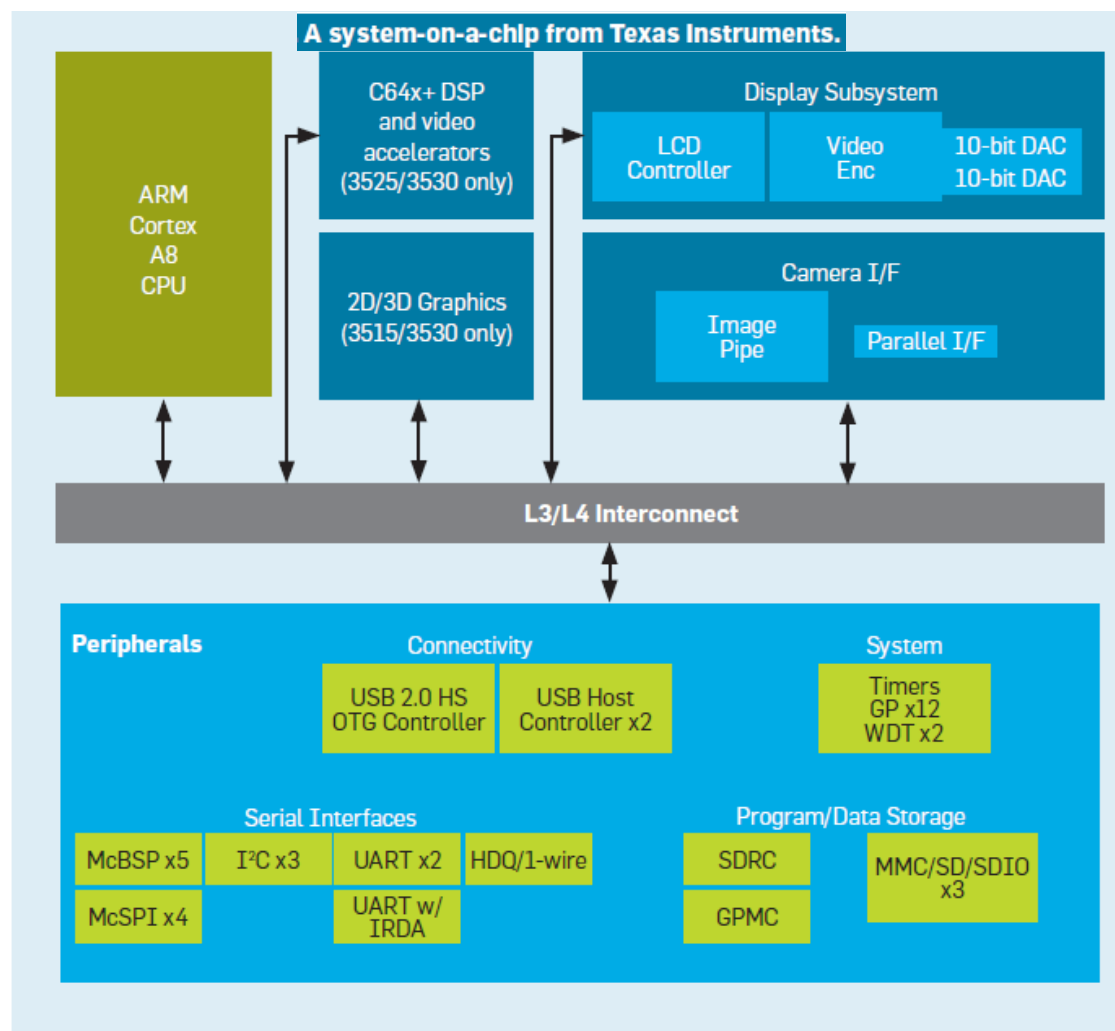
Power and performance of accelerators varying in degree of customization (scaled to 45nm). Customization yields up to 1,000 - 10,000x vs mainstream processors such as Sandybridge.

Source: Apala Guha et al., Systematic Evaluation of Workload Clustering for Designing 10x10 Architectures, ACM SIGARCH Computer Architecture News, vol 41, no2, pp22-29, 2013, <http://dl.acm.org/citation.cfm?id=2490307>

Heterogeneous Architectures

Imagine the impact...

TI's KeyStone SoC + HP Moonshot



"TI's KeyStone II-based SoCs, which integrate fixed- and floating-point DSP cores with multiple ARM® Cortex™A-15 MPCore processors, packet and security processing, and high speed interconnect...."



HP Project Moonshot is dedicated to designing extreme low-energy server technologies

We are pursuing HPC cartridges with HP and TI

Source: S. Borkar, A. Chien, The Future of Microprocessors

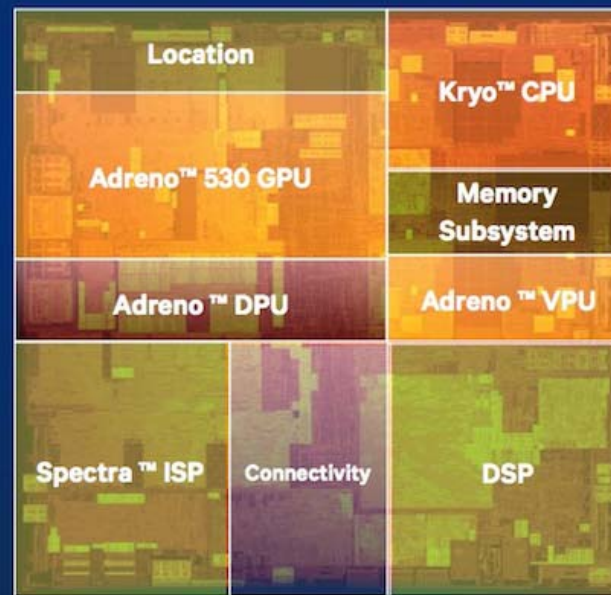
<http://cacm.acm.org/magazines/2011/5/107702-the-future-of-microprocessors/fulltext>

UNIVERSITY of HOUSTON

Heterogeneous Architectures

Snapdragon 820 Overview

Heterogeneous design for high performance with optimized power and thermals



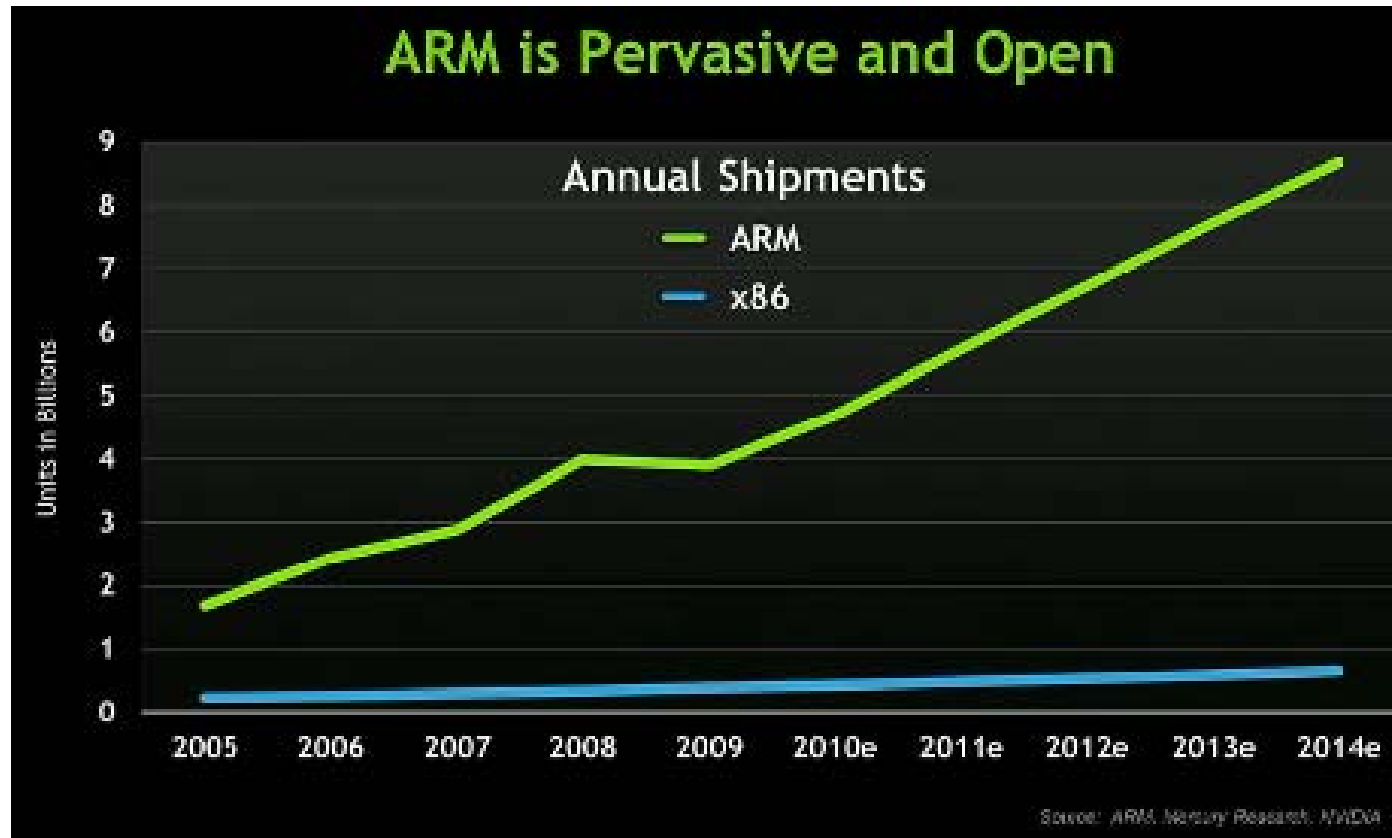
* Not to scale

Snapdragon 820 Mobile SoC

Qualcomm Spectra, Kryo and Adreno are products of Qualcomm Technologies, Inc.

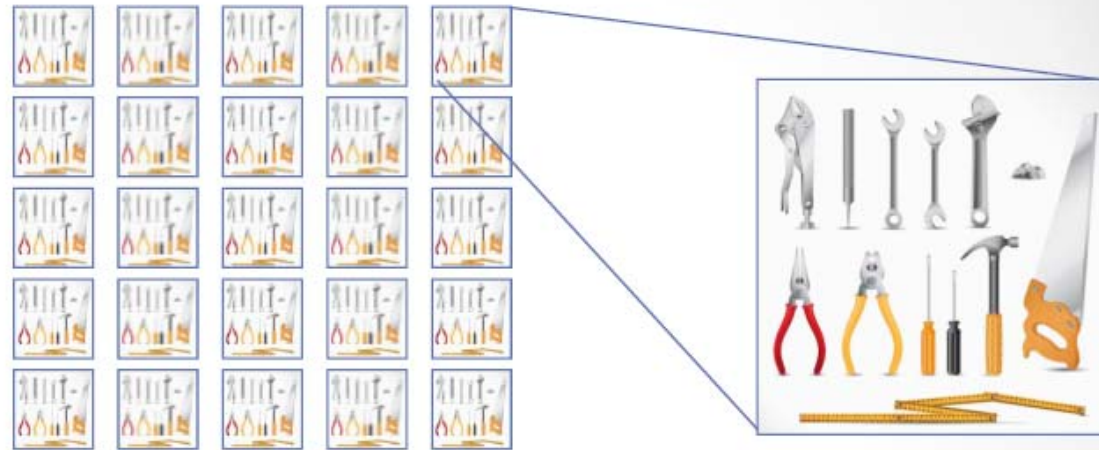
<http://arstechnica.com/gadgets/2015/08/snapdragon-820-is-official-a-look-at-its-gpu-and-how-much-the-chip-matters/>

An aside – Big vs. Small



January 06, 2011. On Wednesday, the GPU-maker -- and soon to be CPU-maker -- revealed its plans to build heterogeneous processors, which will encompass high performance ARM CPU cores alongside GPU cores. The strategy parallel's AMD's Fusion architectural approach that marries x86 CPUs with ATI GPUs on-chip.

10x10 Microprocessor Architecture



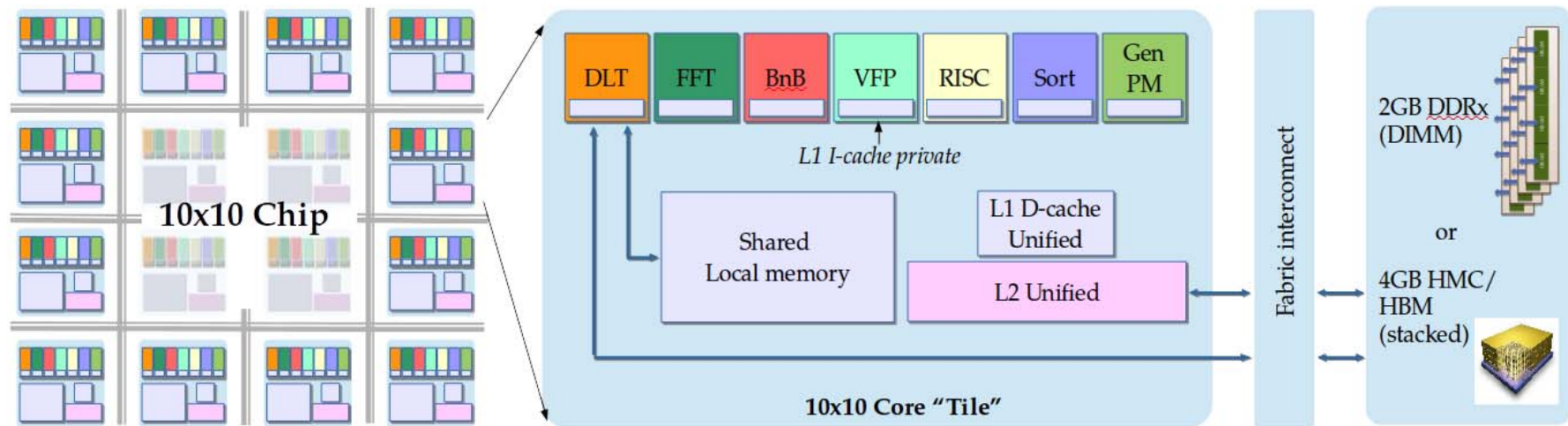
- Many cores, each is 10 distinct accelerators achieving 100x better energy efficiency
 - 10x lower power enables 10x more cores
 - 10x better application performance on a core delivers 100x better overall performance
- Energy is the key limit, 10x10 approaches will outperform traditional
 - Produce highest performance per “core” (optimized implementation)
 - Produce highest performance chip (lowest energy/ops)

Source: A. Chien, <http://www.lanl.gov/conferences/salishan/salishan2010/pdfs/Andrew%20A.%20Chien.pdf>

UNIVERSITY of HOUSTON

10x10 Assessment - Architecture

7 Micro-engines
(not 10 in current assessment)



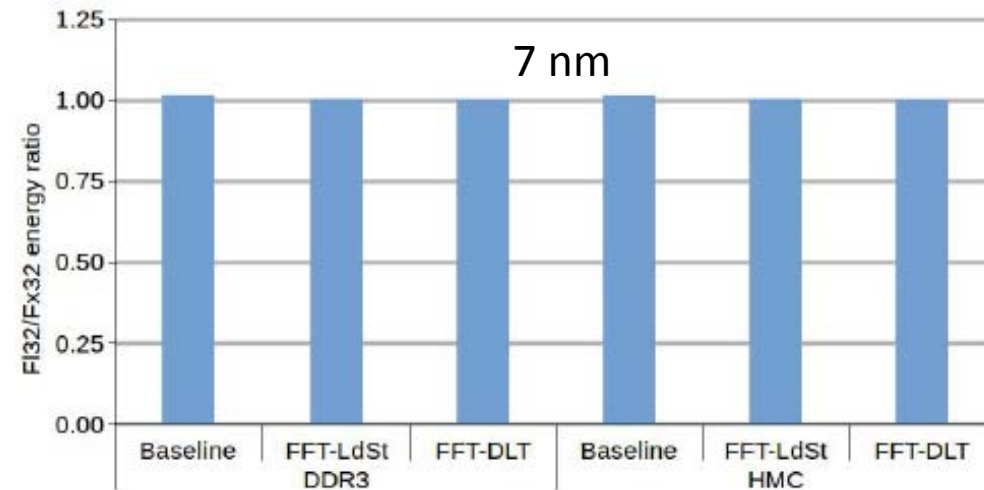
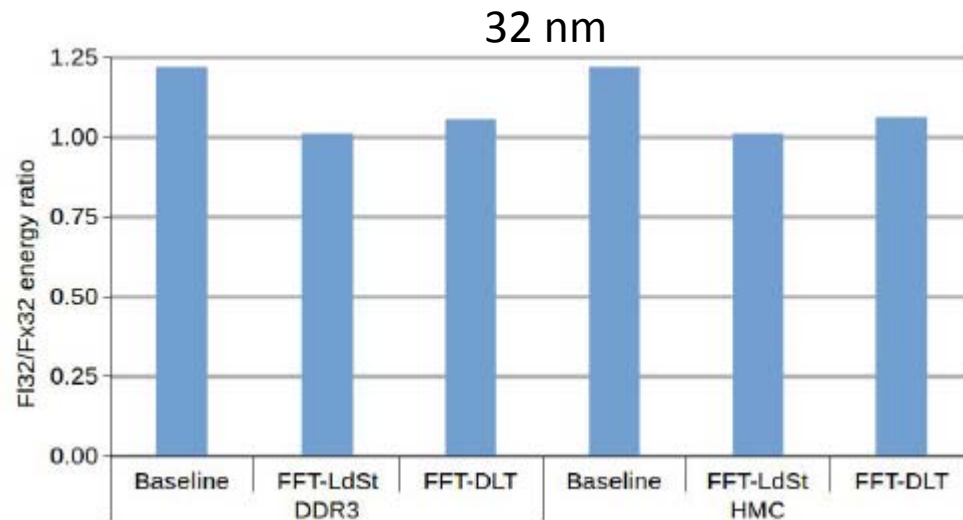
- One RISC core, MIPS-like instruction set, 5-stage pipeline
- Six specialized micro-engines

Source: A. Chien et. al. "10x10 A Case Study in Federated Heterogeneous Architecture", Dept. of Comp. Science, U. Chicago, January 2015



Relative Core Energy: Floating-point vs Integer

FFT-LdSt: impact of
FFT μ -engine
FFT-DLT: impact of FFT
and DLT μ -engines

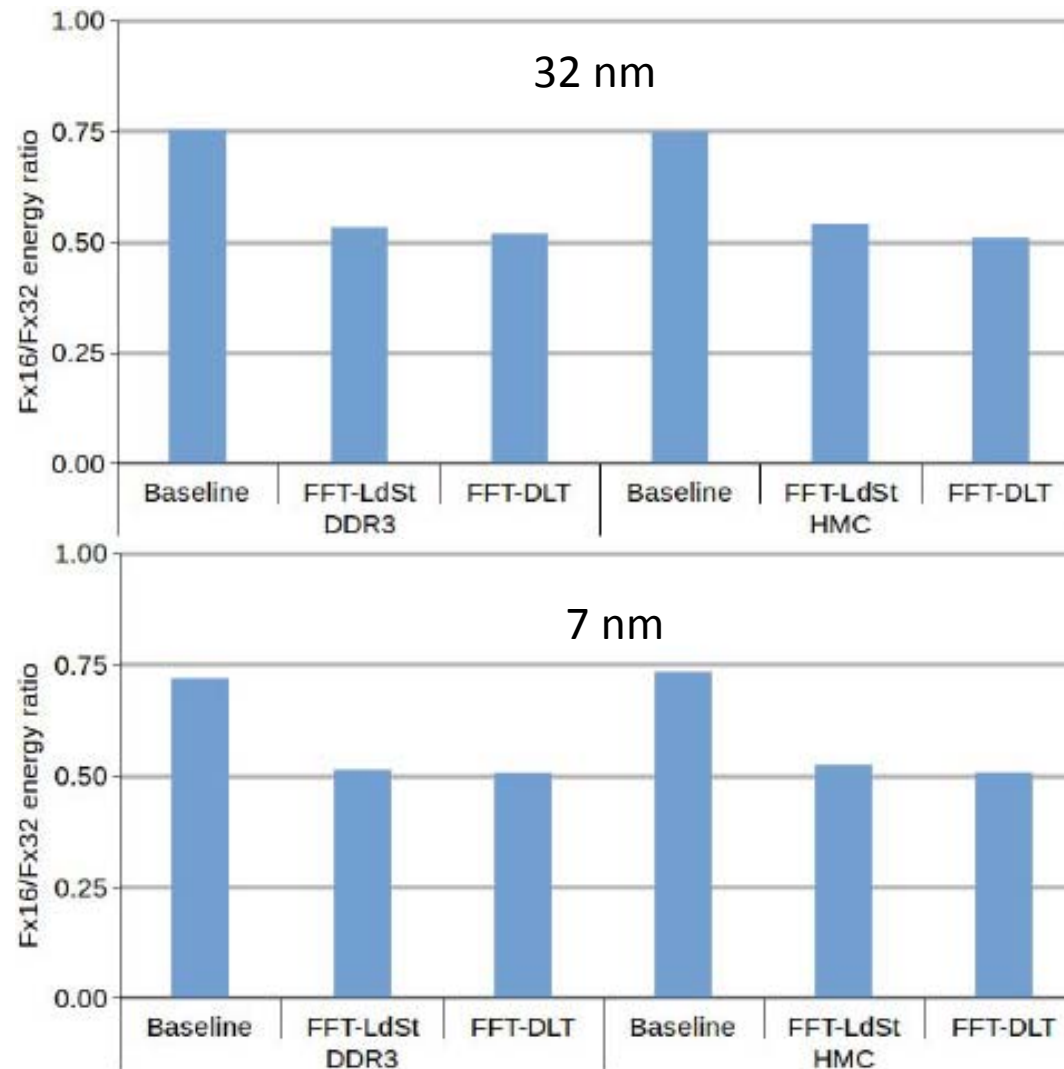


Conclusion: At 7 nm energy
consumption for
computation is practically
independent of integer vs
floating-point format

Source: T. Thank-Hoang et al., Does Arithmetic Logic Dominate Data Movement? A Systematic Comparison of Energy-Efficiency for FFT Accelerators, TR-2015-01, U. Chicago, March 2015



Relative System Energy 16-bit vs 32-bit



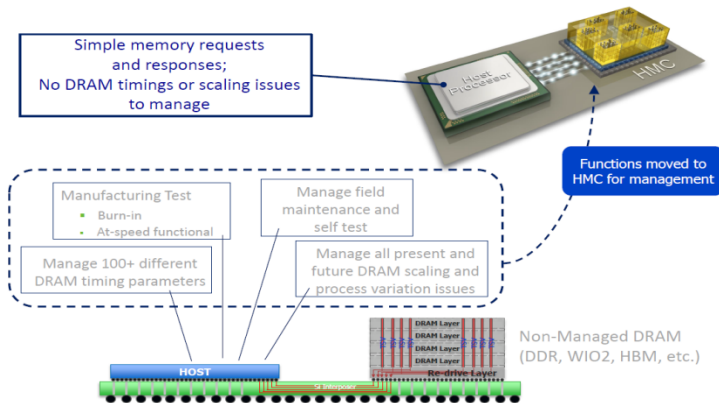
Source: T. Thank-Hoang et al., Does Arithmetic Logic Dominate Data Movement? A Systematic Comparison of Energy-Efficiency for FFT Accelerators, TR-2015-01, U. Chicago, March 2015



Lennart Johnsson
2015-02-06

The Hybrid Memory Cube

Simple HMC Memory Management



HMC_{Gen1}: Technology Comparison

Generation 1 (4 + 1 memory configuration)

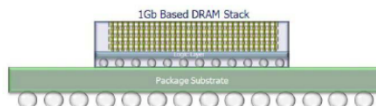
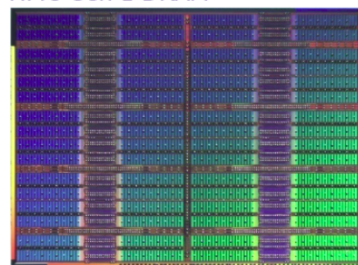
Technology	VDD	IDD	BW GB/s	Power (W)	mW/GB/s	pJ/bit	real pJ/bit
SDRAM PC133 1GB Module	3.3	1.50	1.06	4.96	4664.97	583.12	762
DDR-333 1GB Module	2.5	2.19	2.66	5.48	2057.06	257.13	245
DDR2-667 2GB Module	1.8	2.88	5.34	5.18	971.51	121.44	139
DDR3-1333 2GB Module	1.5	3.68	10.66	5.52	517.63	64.70	52
DDR4-2667 4GB Module	1.2	5.50	21.34	6.60	309.34	38.67	39
HMC, 4 DRAM w/ Logic	1.2	9.23	128.00	11.08	86.53	10.82	13.7

Simple calculation from IDD7 (SDRAM IDD4)

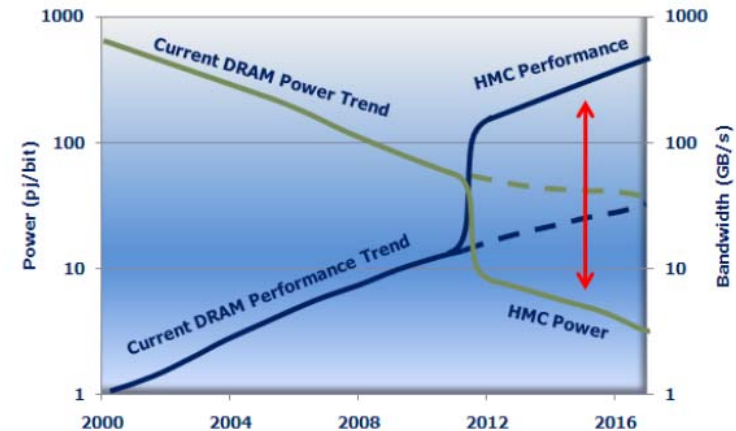
Real system,
some with lower
density modules

- 1Gb 50nm DRAM Array
- 90nm prototype logic
- 512MB total DRAM cube
- 128GB/s Bandwidth
- 27mm x 27mm prototype
- Functional demonstrations!
- Reduced host CPU energy

HMC Gen 1 DRAM



Source: J. Thomas Pawlowski, Micron,
HotChips23, August 17 – 19, 2011



Requirements

Channel Complexity

90% simpler than DDR3L
88% simpler than DDR4

Board Footprint

95% smaller than DDR3L
94% smaller than DDR4

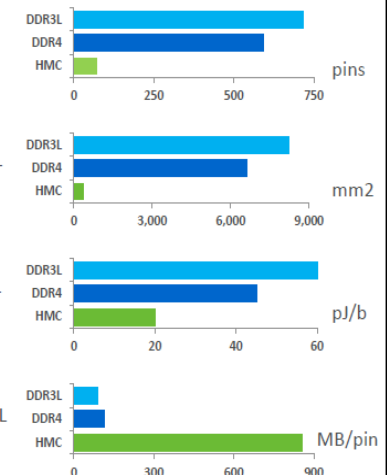
Energy Efficiency

66% greener than DDR3L
55% greener than DDR4

Bandwidth

10.2X greater than DDR3L
8.5X greater than DDR4

TCO Valuation



Source: Todd Farrell, <http://www.lanl.gov/conferences/salishan/salishan2014/Farrell.pdf>

UNIVERSITY of HOUSTON

WSOU-ARISTA001530



Lennart Johnsson
2015-02-06

Processor/Platform Power Management



Fine-grain Power Management—Many Core

Mode		Power Saving	Wake up
Normal	All active	-	-
Standby	Logic off Memory on	50%	Fast
Sleep	Logic and Memory off	80%	Slow

Dynamic, within a core
21 sleep regions per tile (not all shown)

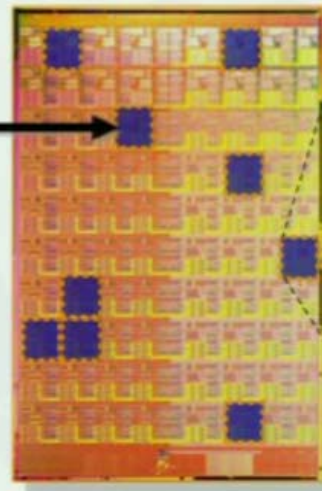
Dynamic Chip Level

STANDBY:

- Memory retains data
- 50% less power/tile

FULL SLEEP:

- Memories fully off
- 80% less power/tile



Data Memory

Sleeping:
57% less power

FP Engine 1

Sleeping:
90% less power

Instruction Memory

Sleeping:
56% less power

FP Engine 2

Sleeping:
90% less power

Router

Sleeping:
10% less power
(stays on to pass traffic)

Energy efficiency increases by 60%,

(from 12 GF/W without to 19.4 GF/W with fine-grain power management)

Vangal et al, An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS, JSSC, January 2008

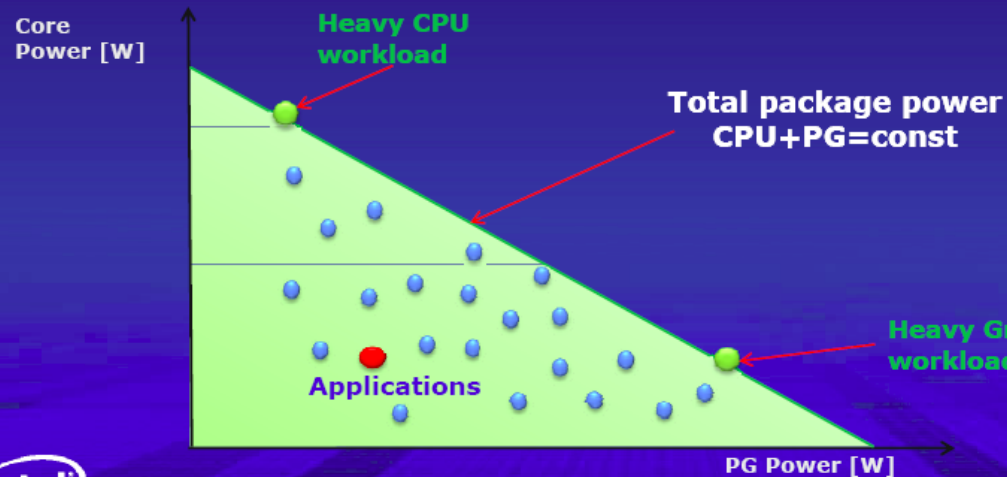
http://isscc.org/media/2012/plenary/David_Perlmutter/SilverlightLoader.html

UNIVERSITY of HOUSTON

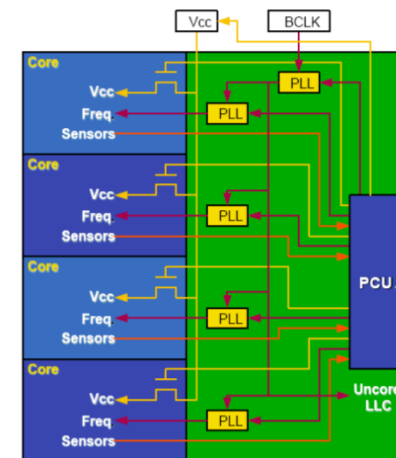
Intel Sandy Bridge CPU

Intel® Turbo Boost Technology 2.0 - Package

- Power specification is defined for the entire package
 - Monolithic die – power budget shared by CPU and PG
 - Sum of component power at or below specifications



Power Control Unit



Integrated proprietary microcontroller
Shifts control from hardware to embedded firmware
Real time sensors for temperature, current, power
Flexibility enables sophisticated algorithms, tuned for current operating conditions

Source: Key Nehalem Choices, Glenn Hinton, Intel,
<http://www.stanford.edu/class/ee380/Abstracts/100217-slides.pdf>

Source: Efi Rotem, Alon Naveh, Doron Rajwan, Avinash Ananthakrishnan, Eli Weissmann
<http://www.hotchips.org/hc23> 2011-08-17 -- 19



Software

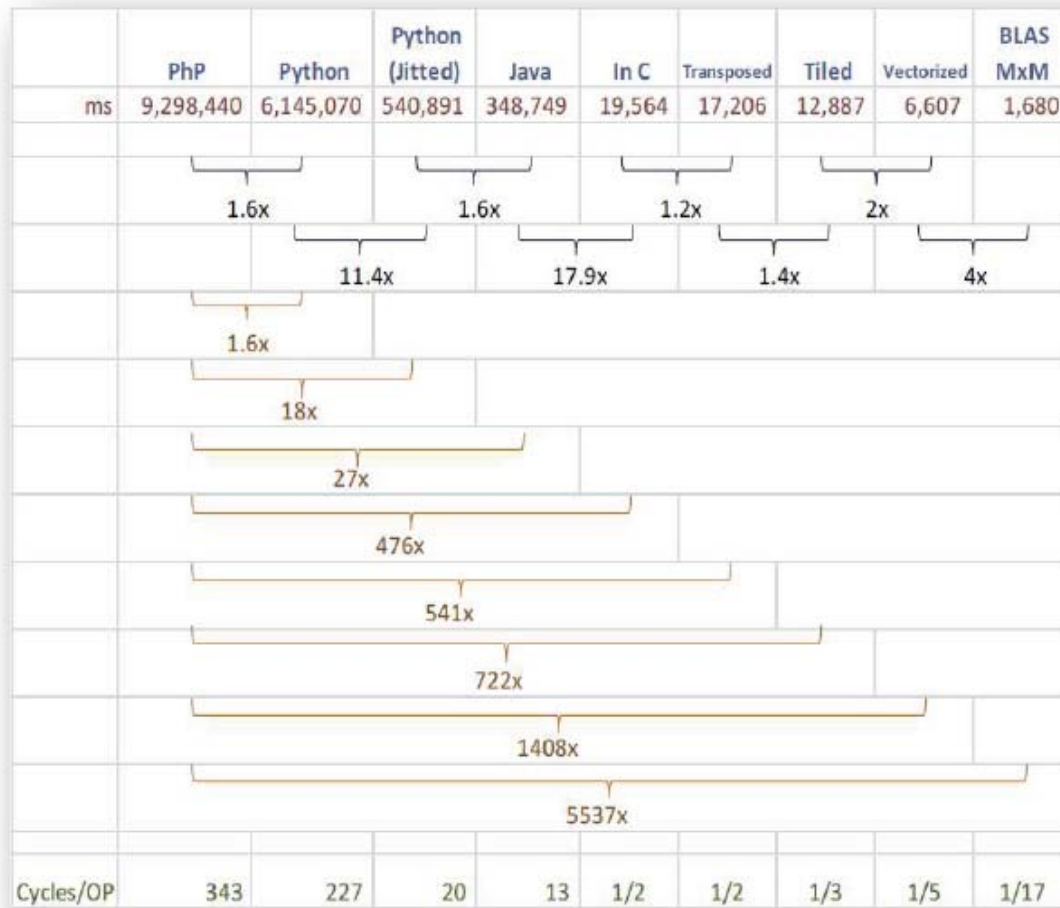
Challenge	Near-Term	Long-Term
1,000-fold software parallelism	Data parallel languages and "mapping" of operators, library and tool-based approaches	New high-level languages, compositional and deterministic frameworks
Energy-efficient data movement and locality	Manual control, profiling, maturing to automated techniques (auto-tuning, optimization)	New algorithms, languages, program analysis, runtime, and hardware techniques
Energy management	Automatic fine-grain hardware management	Self-aware runtime and application-level techniques that exploit architecture features for visibility and control
Resilience	Algorithmic, application-software approaches, adaptive checking and recovery	New hardware-software partnerships that minimize checking and recomputation energy

Source: S. Borkar, A. Chien, The Future of Microprocessors
<http://cacm.acm.org/magazines/2011/5/107702-the-future-of-microprocessors/fulltext>



Energy -Software

Matrix Multiplication – Programmer Productivity vs (Energy) Efficiency



Source: Jim Larus, HiPEAC 2015

6.172 Saman Amarasinghe, MIT Fall 2011

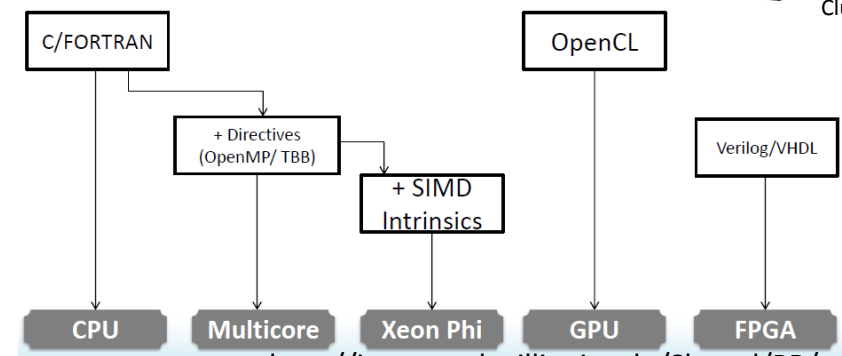
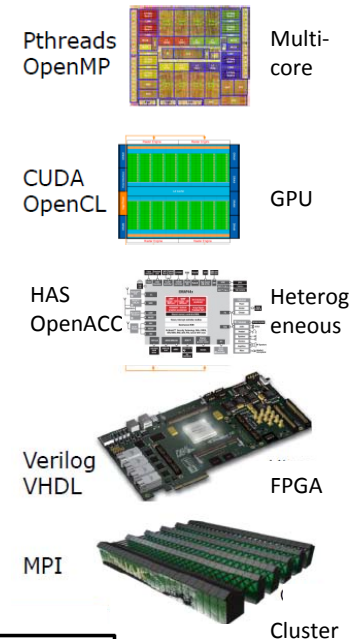


Software

- Explicit memory management – no cache coherence
- Explicit power management (thread/task allocation, voltage and frequency control)
- Complex programming environment – from VHDL (FPGA) to system level, libraries,

Programming Heterogeneous Computer Systems is Very Complex

- ACPI, IPMI, BMC, UEFI,
- Explicit power management (thread/task allocation, voltage and frequency control)
- DMA, Assembly,
- Explicit memory management – no cache coherence
- Linux, Embedded Linux,
- Core: C, Embedded C, Fortran, SIMD,
- Multicore: OpenMP, Pthreads, TBB, Cilk,
- Heterogeneous: OpenCL, OpenACC, HAS, BOLT, ...
- FPGA: Verilog, VHDL
- Clusters: MPI, UPC, Titanium, X10, Co-Array Fortran, Global Arrays, ...



<http://impact.crhc.illinois.edu/Shared/PR/Distinguished-lecture-U-Chicagp-1-22-2015.pdf>



Lennart Johnsson
2015-02-06

Thank You!